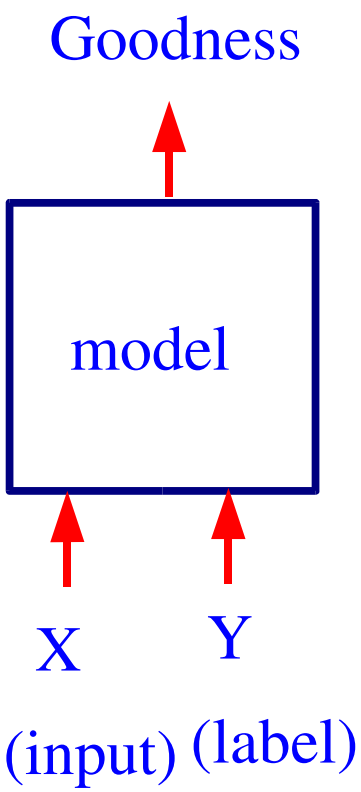
Energy-Based Models

Yann LeCun

The Courant Institute of Mathematical Sciences

New York University

A Model is Designed and trained to Answer Questions



- **Example:** X is an image from a camera; Y is a discrete variable e.g. Y in {animal, human, plane, truck, car}.
- 1. **Best Guess for Y:** which category best describes X?
- 2. **Ranking on Y:** Is X more a car than an airplane?
- 3. **Distribution on Y:** give an estimate of P(animal | X)
- 4. **Best Guess for X:** among all images of airplane, give me the best one.
- 5. **Ranking on X:** is this image more of a truck than that one?
- 6. **Distribution on X:** among all images of airplanes, how likely is this one?

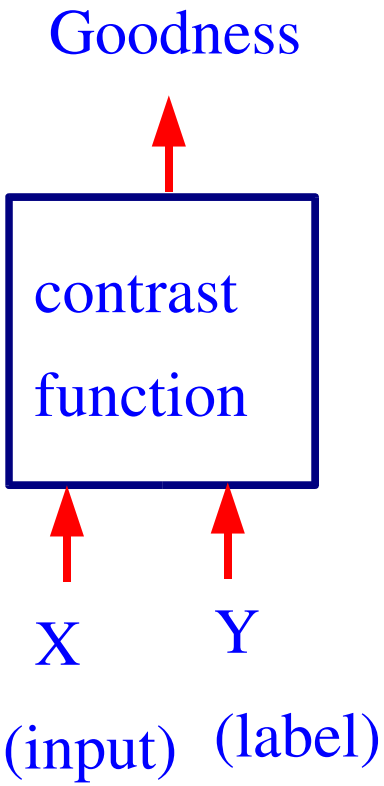
For each question, a different learning strategy is required

Do not answer a more complex question than necessary

- 1. Best Guess for Y: which category best describes X?
- 2. Ranking on Y: Is X more a car than an airplane?
- 3. Distribution on Y: give an estimate of P(animal | X)
- 4. Best Guess for X: among all images of airplane, give me the best one.
- 5. Ranking on X: is this image more of a truck than that one?
- 6. Distribution on X: among all images of airplanes, how likely is this one?

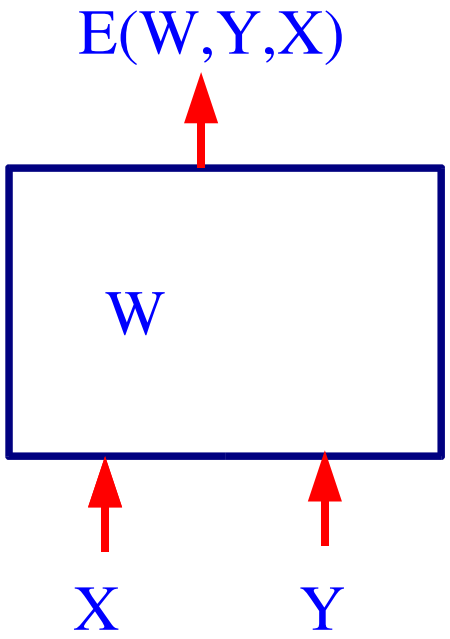
- The questions are in increasing order of complexity.
- The machine should be designed and trained to answer the simplest question possible.

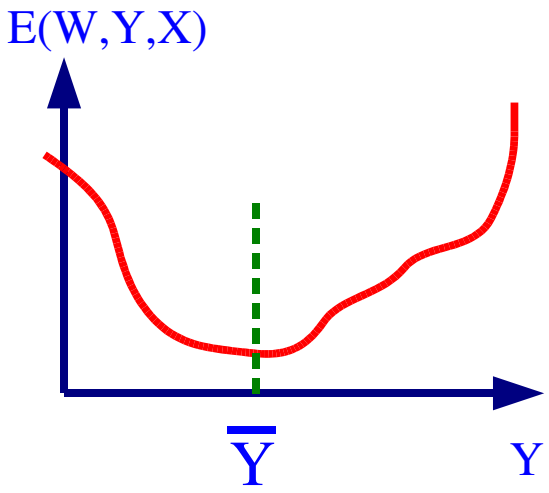
What is a model?



- A model measures the goodness of a combination of observed variables (X) and variables to be predicted (Y).
- Probabilistic approaches compute the distribution P(Y|X), and choose the Y that maximizes it.
- Sometimes, we do not need probabilities.
- Example: driving a robot. When the robot faces an obstacle, it MUST turn left of right. Computing a distribution of steering angles is of little use.
- Question: why estimate the whole distribution P(Y|X) when we are only interested in picking the best value of Y?

Energy-Based Models





- **E(W,Y,X):** is a scalar energy function (a.k.a. Contrast function) that measures the "compatibility" between Y and X.
- **W** is the parameter to be learned.
- **MAP Inference**: Given an input X, find the value of Y that minimizes the energy:

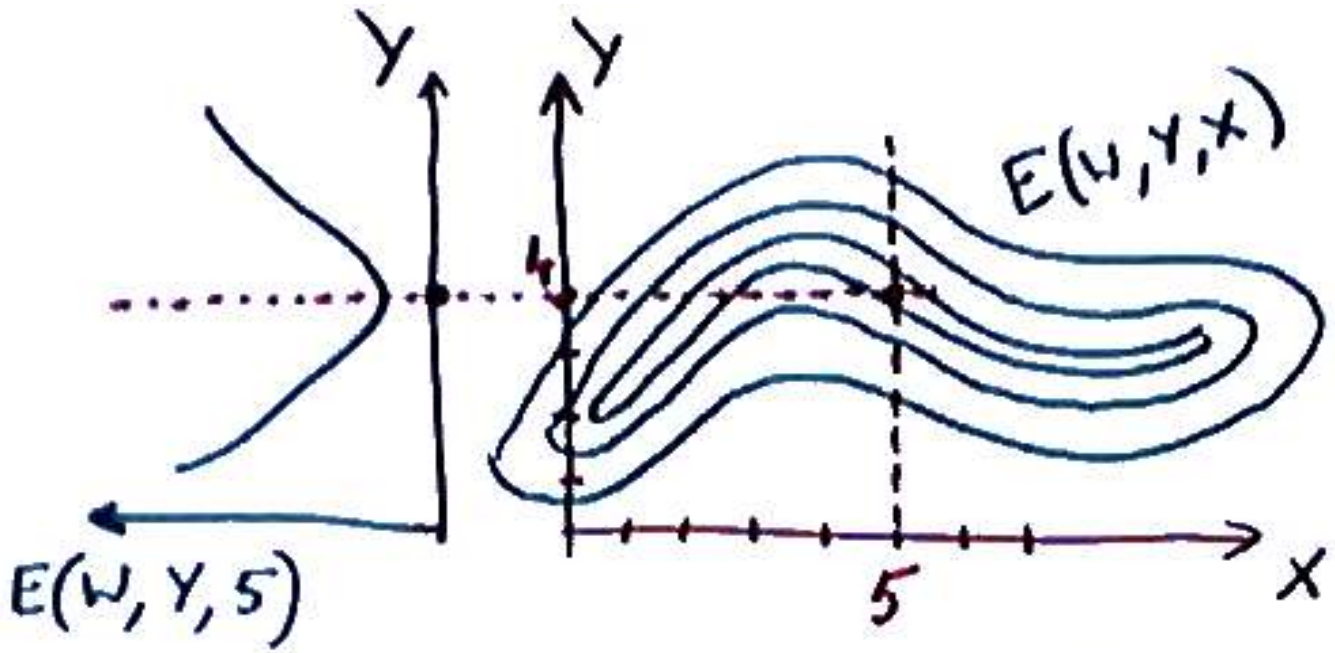
$$\check{Y} = \text{argmin}_{y \in \{Y\}} E(W, y, X)$$

- **Probabilistic Prediction**: Given an input X, compute the conditional distribution over Y (Gibbs Distribution):

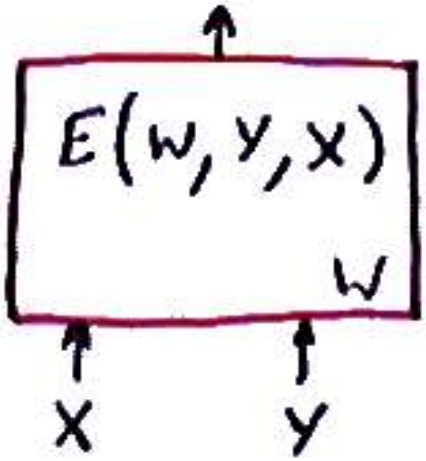
$$P(Y|X) = \frac{\exp\left(-\beta E(W, Y, X)\right)}{\sum_{y \in \{Y\}} \exp\left(-\beta E(W, y, X)\right)}$$

- **For decision making, we need no normalization.**

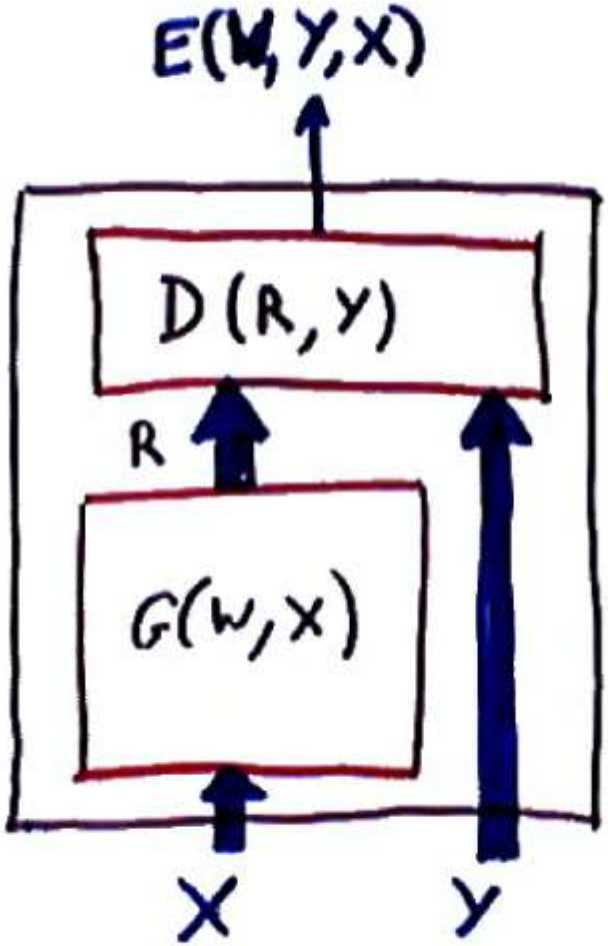
MAP Inference with Energy-Based Models



$$\check{y} = \min_{y} E(w, y, x)$$

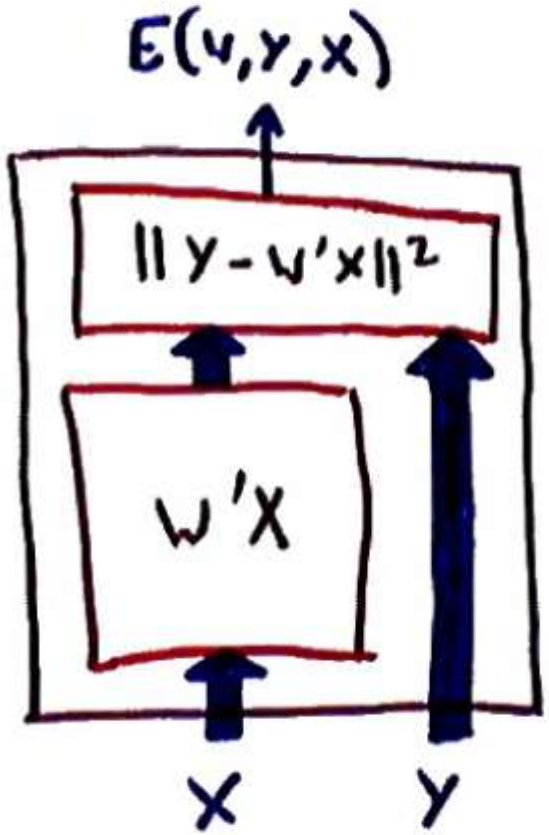


Examples of EBM: Regressor



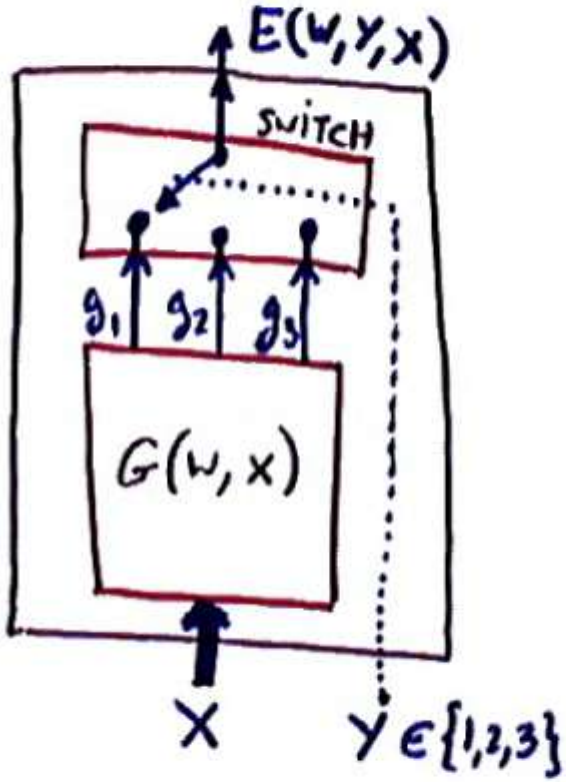
- X and Y are vectors or other entities
- Energy: $E(W, Y, X) = D(Y, G(W, X))$ where $D(Y, R)$ is a distance or dissimilarity measure.
- Best output: $\check{Y} = \min_Y E(W, Y, X) = G(W, X)$.

Examples of EBM Regressor: Linear Regression



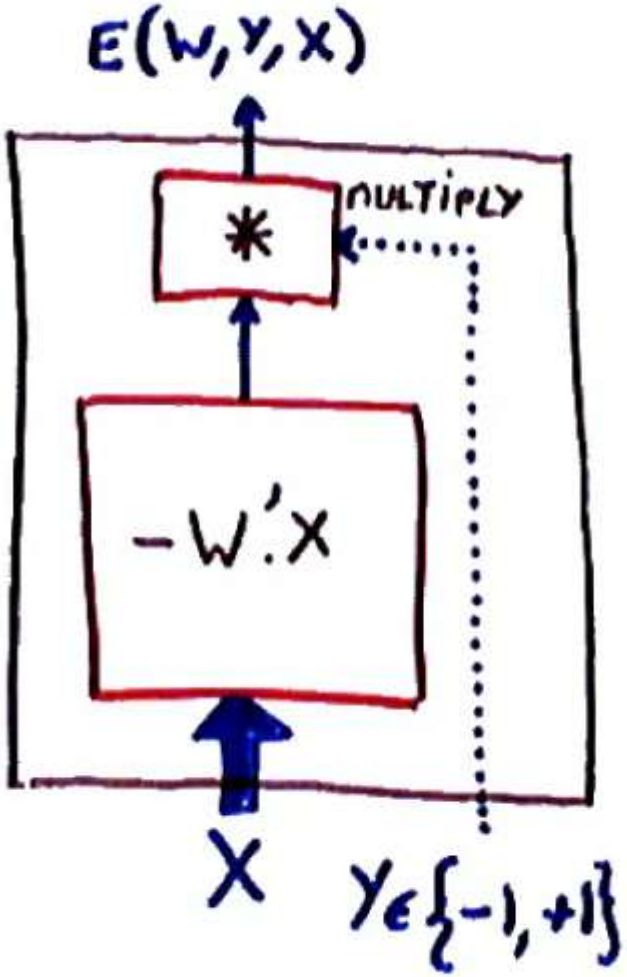
- X and Y are vectors
- Energy: $E(W, Y, X) = ||Y - W'X)||^2$.
- Best output: $\check{Y} = \min_Y E(W, Y, X) = W'X$.

Examples of EBM: Classifier



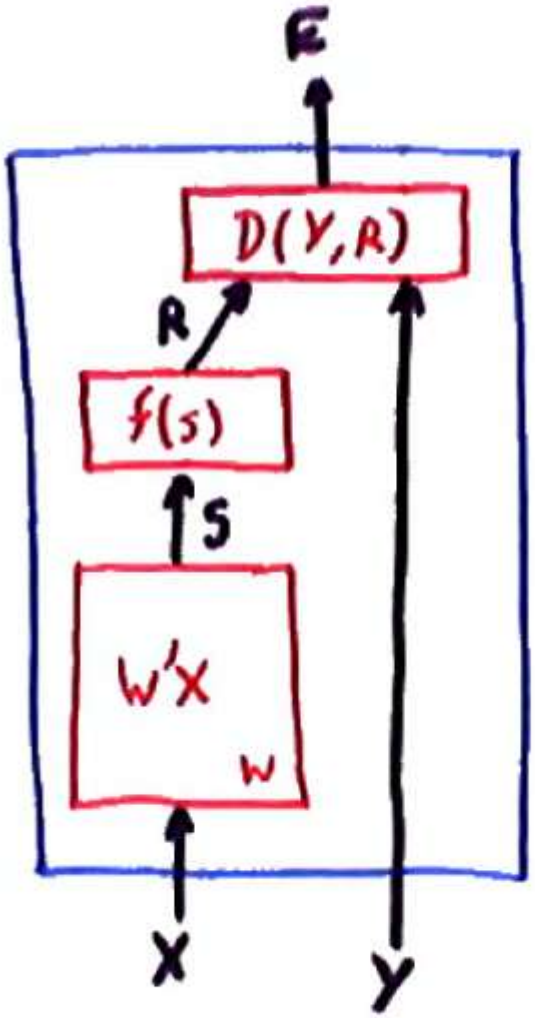
- Y is a discrete variable, $\{Y\} = \{1, 2, 3\}$.
- Energy: $E(W, Y, X) = \sum_k G_k(W, X)\delta(k, Y)$, where $\delta(k, Y) = 1$ iff $k = Y$ and 0 otherwise.
- $G_k(W, X)$, the k-th component of the output vector of $G(W, X)$ is interpreted as the "cost" of classifying X into category k.
- Best output: $\check{Y} = \min_{Y \in \{Y\}} E(W, Y, X) = \min_k G_k(W, X)$.

Examples of EBM Classifier: Perceptron



- Y is a discrete variable, $\{Y\} = \{-1, +1\}$.
- Energy: $E(W, Y, X) = -Y.W'X$.
- Best output: $\check{Y} = \text{sign}(W'X)$, where $\text{sign}(R) = +1$ iff $R > 0$ and -1 otherwise.

Linear Machines

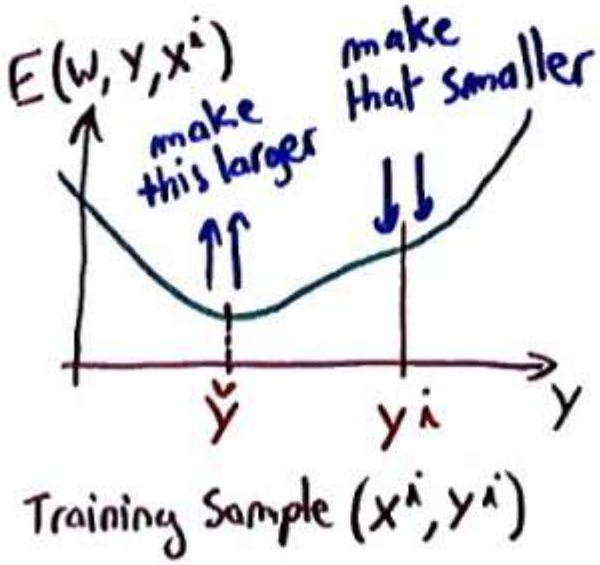


- The learning algorithms we have seen so far (perceptron, linear regression) are of that form, with the assumption that $G(W, X)$ only depends on the dot product of W and X.
- In other words, The E function of 2-class linear classifiers can be written as:

$$E(Y, X, W) = D(Y, f(W'X))$$

  where $W'X$ is the dot product of vectors W and X, and f is a monotonically increasing scalar function.
- in the following, we assume $Y = -1$ for class 1, and $Y = +1$ for class 2.

Training Energy-Based Models



- To train an EBM, we minimize a **loss function**, which is an average over training samples of a **per-sample loss function** $L(W, Y^i, X^i)$:

$$\mathcal{L}(W, \mathcal{S}) = \frac{1}{P} \sum_{i=1}^{P} L(W, Y^i, X^i)$$

- The loss function must be designed so that minimizing it with respect to W will make the machine approach the desired behavior.

To ensure this, we pick loss functions that, for a given training input X^i, will drive the energies $E(W, Y^i, X^i)$ associated with the desired output Y^i to be lower than the energies associated with all other (undesired) outputs values $E(W, Y, X^i)$ for all $Y \neq Y^i, Y \in \{Y\}$.

Form of the Loss Function

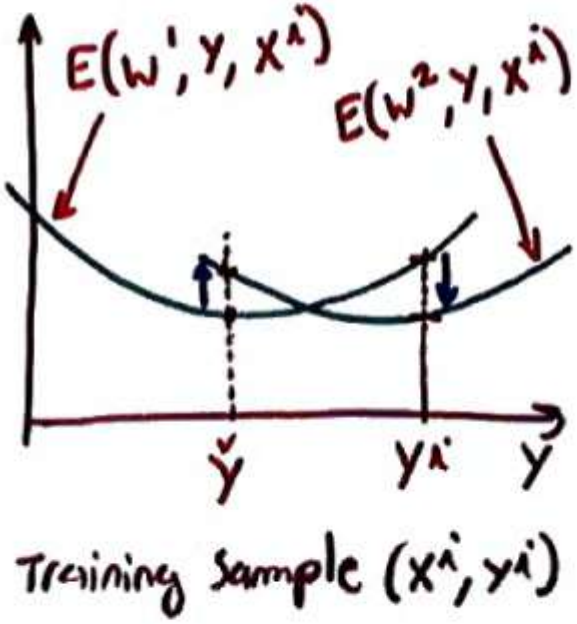
- We assume that the per-sample loss function $L(W, Y^i, X^i)$ has a lower bound over W for all Y^i, X^i.
- We assume that L depends on X^i only indirectly through the set of energies $\{E(W, Y, X^i)\ ,\ Y \in \{Y\}\}$.
- For example, if $\{Y\}$is the set of integers between 0 and $k-1$ (as would be the case for a classifier with k categories), the per-sample loss for sample (X^i, Y^i) should be of the form:

$$L(W, Y^i, X^i) = L(Y^i, E(W, 0, X^i), E(W, 1, X^i), \ldots, E(W, k-1, X^i))$$

- With this assumption, we separate the choice of the loss function from the details of the internal structure of the machine, and limit the discussion to how minimizing the loss function affects the energies.

Examples of Loss: Energy Loss

Energy Loss, the simplest of all losses: $L_{\text{energy}}(W, Y^i, X^i) = E(W, Y^i, X^i)$. This loss only works if $E(W, Y, X^i)$ has a special form which guarantess that making $E(W, Y^i, X^i)$ lower will automatically make $E(W, Y, X^i)$ for $Y \neq Y^i$ larger than the minimum.

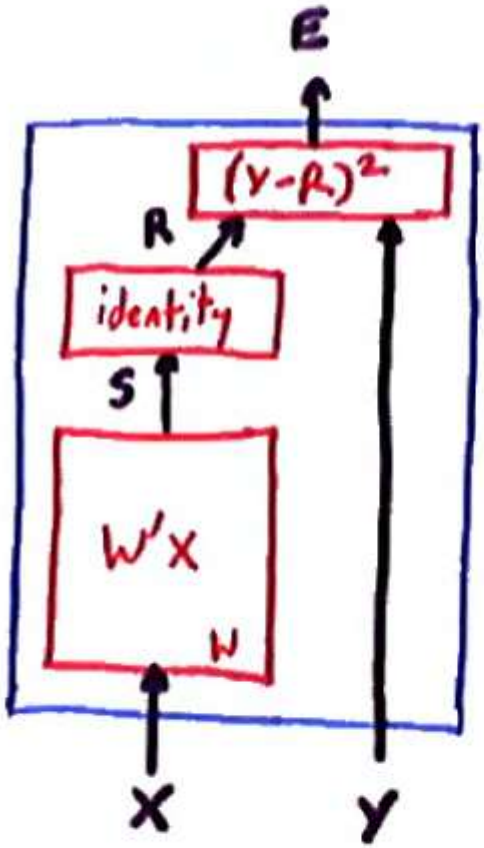


Example: if $E(W, Y, X)$ is quadratic in Y, as is the case for regression with squared error: $E(W, Y, X) = ||Y - G(W, X)||^2$,
Let $W(1)$ is the parameter before a learning update, and $W(2)$ the parameter after the learning update, and let $\check{Y} = \min_Y E(W(1), Y, X)$. Then,

$$E(W(2), Y^i, X^i) - E(W(2), \check{Y}, X^i) < E(W(1), Y^i, X^i) - E(W(1), \check{Y}, X^i)$$

Linear Regression

Linear regression uses the Energy loss

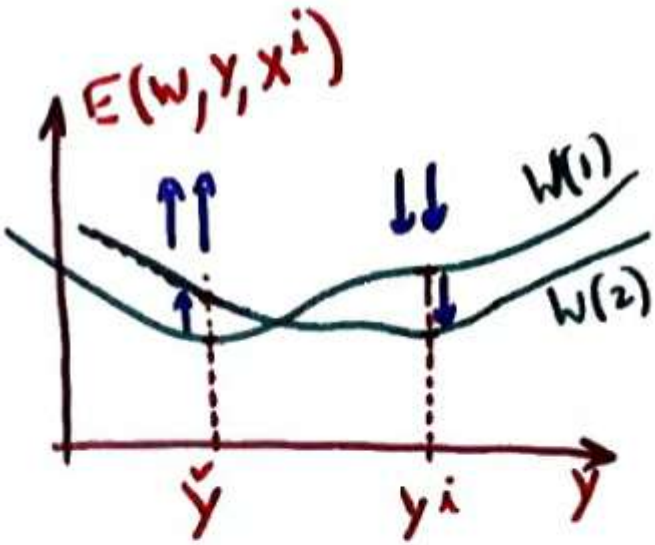


- $R = W'X$
- $E(W, Y, X) = D(Y, R) = \frac{1}{2}||Y - R||^2$
- $L(W, Y^i, X^i) = D(Y^i, W'X^i)$
- $\frac{\partial L}{\partial W} = \frac{\partial D(Y^i,R)}{\partial R} \frac{\partial R}{\partial W}$
- $\frac{\partial L}{\partial W} = \frac{\partial D(Y^i,R)}{\partial R} \frac{\partial (W'X^i)}{\partial W} = (R - Y^i)X^i$
- descent: $W \leftarrow W + \eta(Y^i - R)X^i$

Examples of Loss: Perceptron Loss

Perceptron Loss:

$$L_{\text{perceptron}}(W, Y^i, X^i) = E(W, Y^i, X^i) - \min_{Y \in \{Y\}} E(W, Y, X^i)$$

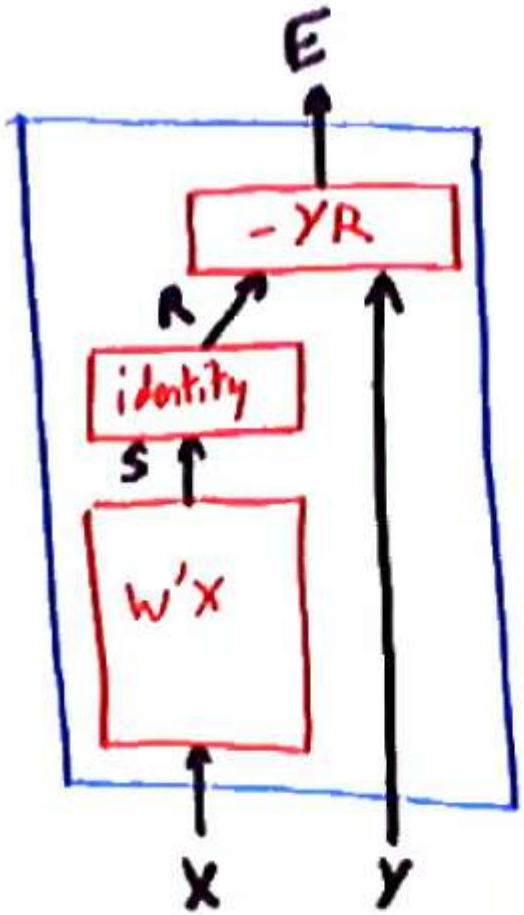


Adjust W so that $E(W, Y^i, X^i)$ gets smaller, while $\check{Y} = \min_{Y \in \{Y\}} E(W, Y, X^i)$ gets bigger (or more precisely, so that the difference decreases).
This algorithm makes no update whenever the energy of the desired Y is lower than all the others.

Perceptron

$$L_{\text{perceptron}}(W, Y^i, X^i) = E(W, Y^i, X^i) - \min_{Y \in \{Y\}} E(W, Y, X^i)$$

$\{Y\} = \{-1, +1\}$.



- $R = W'X$
- $E(Y, X, W) = D(Y, R) = -YR$
- $Y \in \{-1, +1\}$, hence $\min_Y -YR = -\text{sign}(R)R$ where $\text{sign}(R) = 1$ iff $R > 0$, and -1 otherwise.
- $L(W, Y^i, X^i) = -(Y^i - \text{sign}(R))R$
- $\frac{\partial L}{\partial W} = \frac{\partial -(Y^i - \text{sign}(R))R}{\partial R} \frac{\partial R}{\partial W}$
- $\frac{\partial L}{\partial W} = -(Y^i - \text{sign}(W'X^i))X^i$
- descent: $W \leftarrow W + \eta(Y^i - \text{sign}(W'X^i))X^i$

Examples of Loss: Log-Likelihood Loss

Log-Likelihood Loss:

$$L_{\text{ll}}(W, Y^i, X^i) = E(W, Y^i, X^i) + \frac{1}{\beta} \log \left(\sum_{Y \in \{Y\}} \exp(-\beta E(W, Y, X^i)) \right)$$

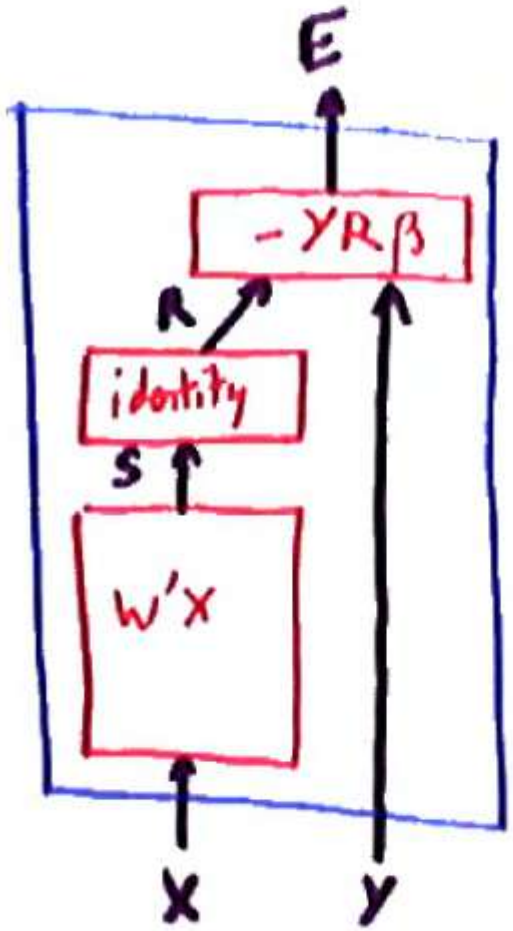
where β is a positive constant.

- The function $\mathcal{F}_\beta(\{Y\}) = \frac{1}{\beta} \log \left(\sum_{Y \in \{Y\}} \exp(-\beta E(W, Y, X^i)) \right)$ is called the **free energy** of the ensemble $\{Y\}$ for temperature $1/\beta$.
- We define $\mathcal{Z}_\beta(\{Y\}) = \sum_{Y \in \{Y\}} \exp(-\beta E(W, Y, X^i))$ as the **partition function** of ensemble $\{Y\}$.
- Interesting property # 1: $\mathcal{F}_\beta(\{Y\}) = \frac{1}{\beta} \log \mathcal{Z}_\beta(\{Y\})$
- Interesting property # 2: $\lim_{\beta \to \infty} \mathcal{F}_\beta(\{Y\}) = \min_{Y \in \{Y\}} E(W, Y, X^i)$

For very large β, the log-likelihood loss reduces to the Perceptron loss.

Logistic Regression (a.k.a MaxEnt)

$$L_{\text{ll}}(W) = E(Y^i, X^i, W) + \log\left(\sum_{Y \in \{Y\}} \exp(-E(W, Y, X^i))\right)$$



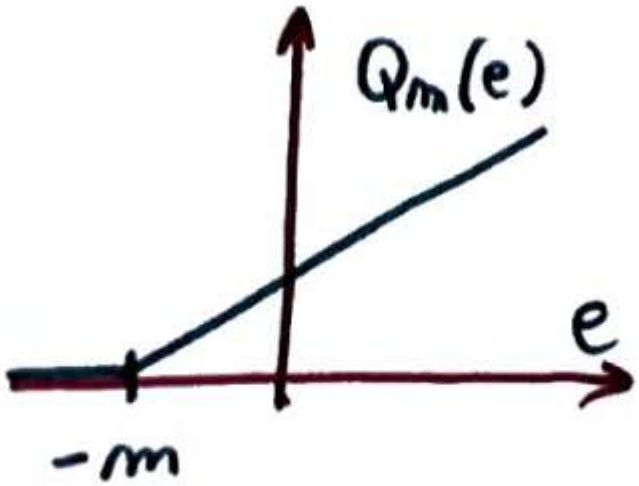
- $R = \frac{1}{2} W'X$
- $E(Y, X, W) = D(Y, R) = -\frac{1}{2} YR = -\frac{1}{2} YW'X$
- $L(W) = \log(1 + \exp(-Y^i W' X^i))$
- $\frac{\partial L}{\partial W} = \frac{\partial D(Y^i, R)}{\partial R} \frac{\partial S}{\partial W}$
- $\frac{\partial L}{\partial W} = -\left(\frac{Y^i + 1}{2} - \frac{1}{1 + \exp(-W' X^i)}\right) X^i$
- descent: $W \leftarrow W + \eta \left(\frac{Y^i + 1}{2} - \frac{1}{1 + \exp(-W' X^i)}\right) X^i$

Examples of Loss: Margin Loss

Margin Loss: for discrete output set $\{Y\}$:

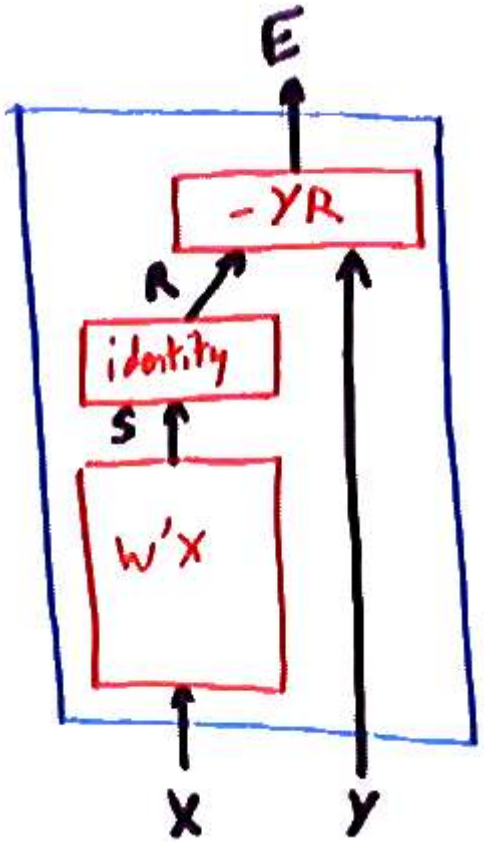
$$L_{\text{margin}}(W, Y^i, X^i) = Q_m\left(E(W, Y^i, X^i) - \min_{Y \in \{Y\}, Y \neq Y^i} E(W, Y, X^i)\right)$$

where $Q_m(e)$ is any function that is monotonically increasing for $e > -m$, where m is a constant called the **margin**.

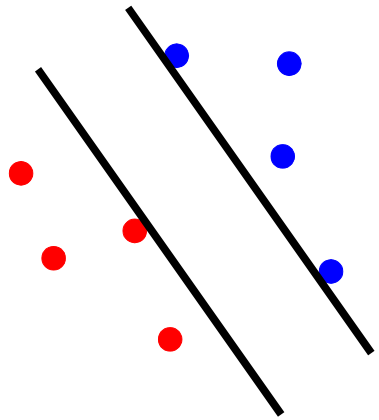


Adjust W so that $E(W, Y^i, X^i)$ gets smaller, while all $E(W, Y, X^i)$ for which $E(W, Y, X^i) - E(W, Y^i, X^i) < m$ get bigger. This guarantees that the energy of the desired Y will be smaller than all other energies by at least m.

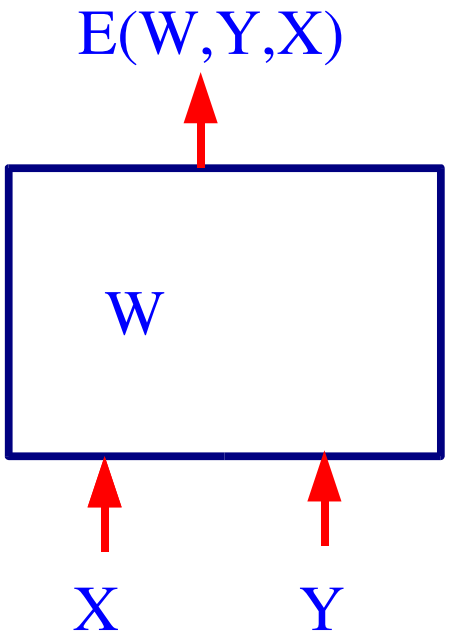
Linear Model + Margin Loss + Regularization = SVM



- Minimize the hinge loss: make the energy of all the "good" answers smaller that the energy of any "bad" answer by at least m (the margin).
- Minimize the Regularization term: Make W as short as possible.
- This is equivalent to keeping ||W|| constant, while maximizing m.

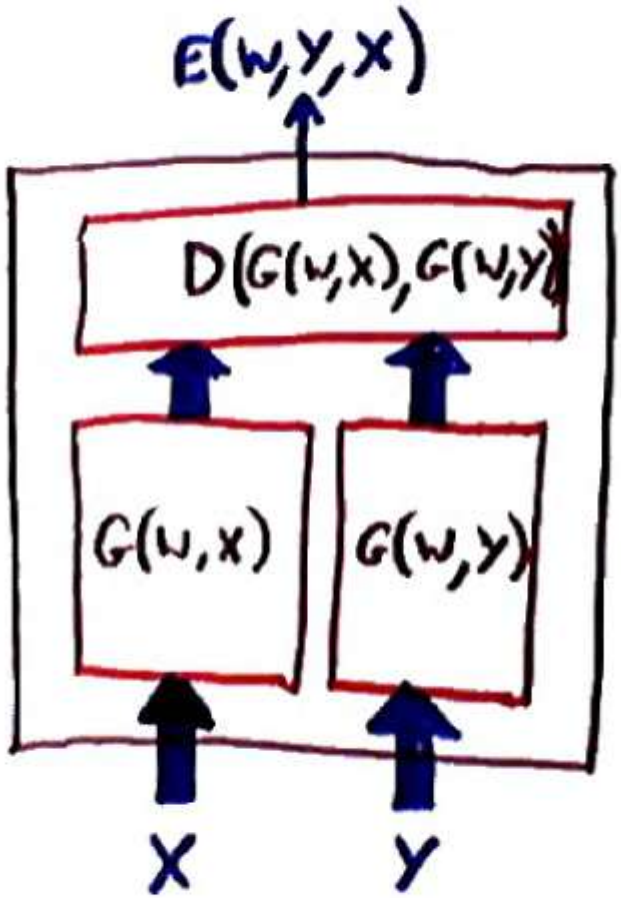


Architecture



- We can put anything we want in the box.
- The energy can be a very complicated non-linear function of X,Y, and W (e.g. A neural net, a graphical model, an HMM, Markov Random Field,....).
- The internal structure of the box is called the architecture of the model.

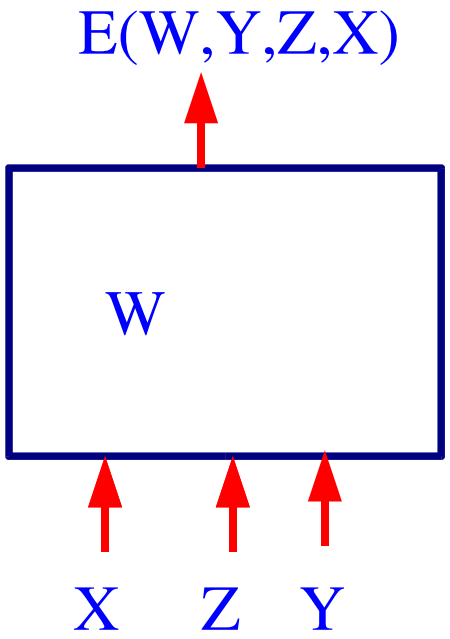
Examples of EBM: Matcher



- X and Y are vectors of the same dimension.
- Energy:
  $E(W, Y, X) = D(G(W, Y), G(W, X))$ where $D(.,.)$ is a distance or dissimilarity measure.
- Best output: $\check{Y} = \min_Y E(W, Y, X) = G^{(} - 1)(G(W, X))$.

Finding the Y that minimizes the energy may be non-trivial

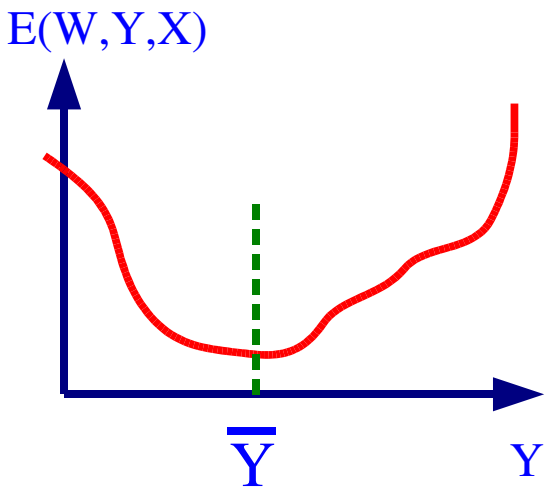
Energy-Based Models with Latent Variables



- **Z:** is the latent variable (never observed).
- **MAP inference:** given an input X, find the value of Y that minimizes the energy:

$$\check{Y} = \text{argmin}_{y \in \{Y\}} \check{E}(W, y, X)$$

$$\check{E}(W, y, X) = \text{argmin}_{z \in \{Z\}} E(W, y, z, X)$$



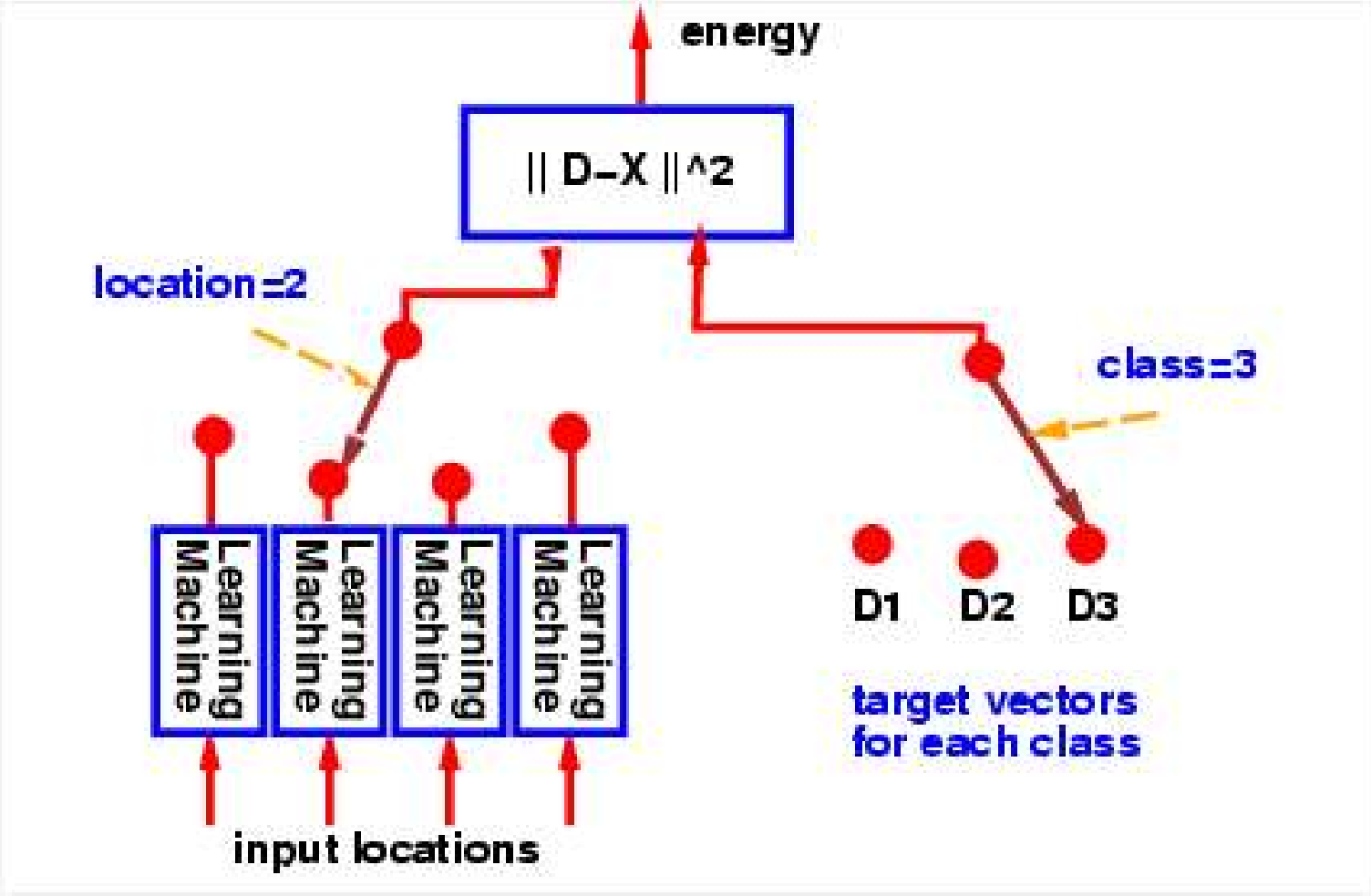
- **Probabilistic Inference**: simply marginalize over Z.

$$P(Y|X) = \frac{\int_{z \in \{Z\}} \exp\left(-\beta E(W, Y, z, X)\right)}{\sum_{y \in \{Y\}} \int_{z \in \{Z\}} \exp\left(-\beta E(W, y, z, X)\right)}$$

What can the latent variables represent?

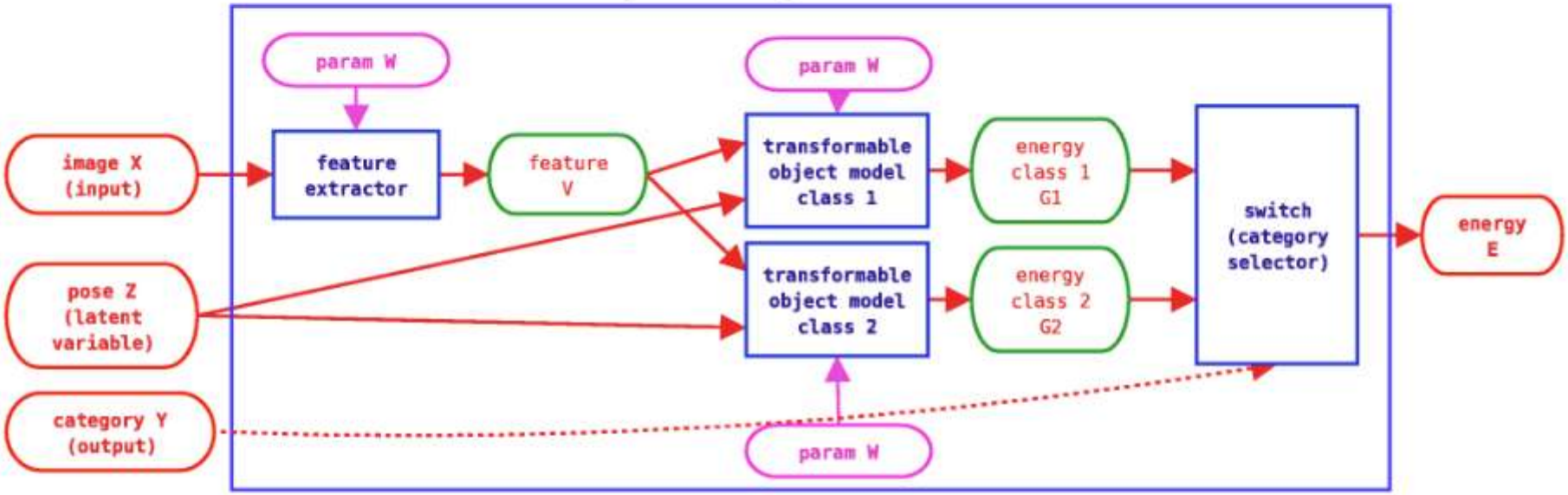
- **Variables that would make the task easier if they were known:**
  - **Face recognition**: the gender of the person, the orientation of the face.
  - **Object recognition**: the pose parameters of the object (location, orientation, scale), the lighting conditions.
  - **Parts of Speech Tagging**: the segmentation of the sentence into syntactic units, the parse tree.
  - **Speech Recognition**: the segmentation of the sentence into phonemes or phones.
  - **Handwriting Recognition**: the segmentation of the line into characters.
- **In general, we will search for the value of the latent variable that allows us to get an answer (Y) of smallest energy.**

Example of latent variable: location



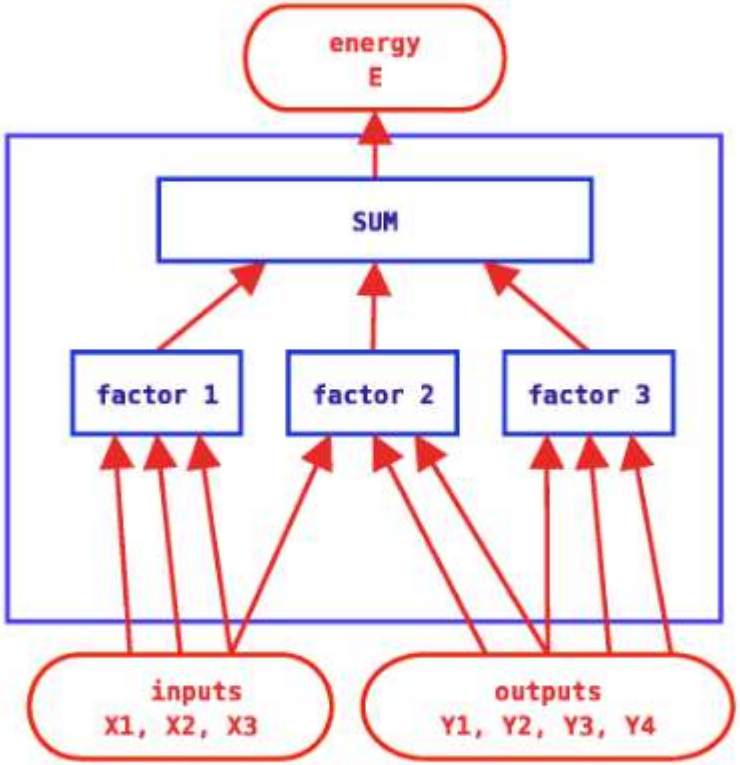
Example of latent variable: pose

EBM Architecture for invariant object recognition



Each object model matches the output of the feature extractor to a reference representation that is transformed by the pose parameters.
Inference finds the category and the pose that minimize the energy.

EBMs as Factor Graphs



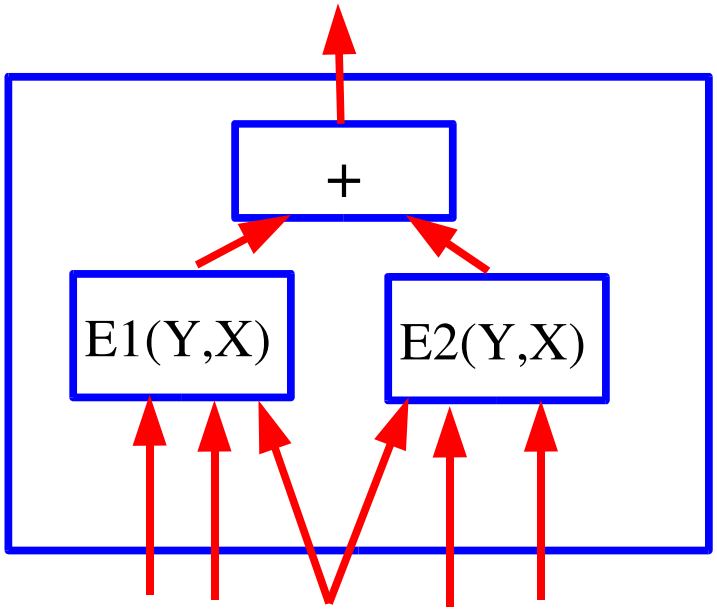
An EBM whose energy function can be "factorized" as a sum of individual functions (factors) is equivalent to a graphical model represented as a factor graph.
Any traditional graphical model can be formulated as a factor graph, but the converse is not true. Each factor is akin to $-log$ of the potential functions of a clique of variable nodes.

Efficient inference algorithms such as (loopy) belief propagation can be used to compute the marginals of Y, or the lowest energy (MAP) configuration [Kschischang, Frey, Loeliger, 2001].

Energy-Based Graphical Models

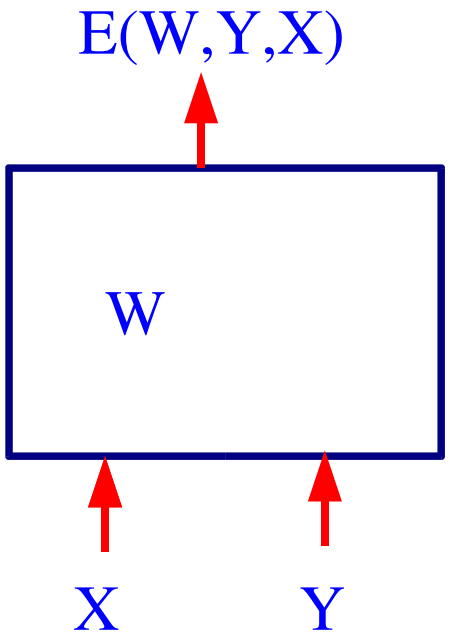
- **A factor graph is a general way to represent a graphical model.**
- **Probabilistic Factor Graph:**

$$P(Y|X) = \frac{\prod_i \Psi^i(Y,X)}{\int_y \prod_i \Psi^i(y,X)}$$



- **Energy-Based Factor Graph:**
  - No latent vars: $P(Y|X) = \frac{\exp(-\beta \sum_i E^i(Y,X))}{\int_y \exp(-\beta \sum_i E^i(y,X))}$
  - Latent vars $P(Y|X) = \frac{\int_z \exp(-\beta \sum_i E^i(Y,z,X))}{\int_{y,z} \exp(-\beta \sum_i E^i(y,z,X))}$

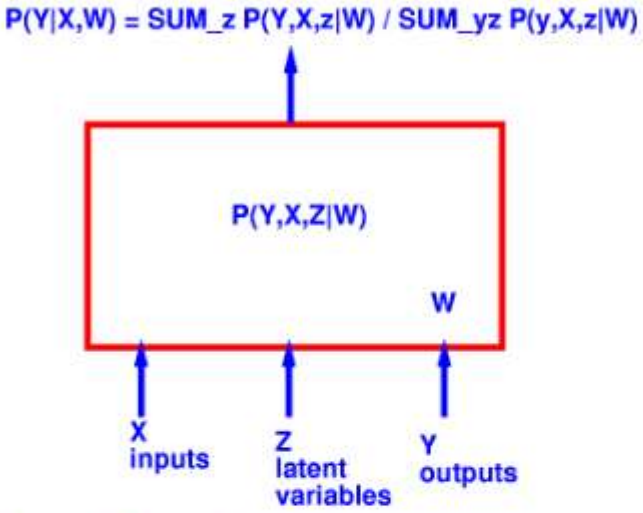
Architecture + Inference Algo + Loss Function = Model



- **1. Design an architecture:** a particular form for E(W,Y,X).
- **2. Pick an inference algorithm for Y:** MAP or conditional distribution, belief prop, min cut, variational methods, gradient descent, MCMC, HMC.....
- **3. Pick a loss function:** in such a way that minimizing it with respect to W over a training set will make the inference algorithm find the correct Y for a given X.
- **4. Pick an optimization method.**

- **PROBLEM: What loss functions will make the machine approach the desired behavior?**

Training Probabilistic Models



Training set: $\mathcal{S} = \{(X^1, Y^1), ...(X^p, Y^p)\}$.
Training Criterion: Max Likelihood

$$\prod_{i=1}^{p} P(Y^i|X^i, W) = \prod_{i=1}^{p} \frac{\int_z P(W, Y^i, z, X^i)}{\int_{yz} P(W, y, z, X^i)}$$

Loss Function: Negative Log Likelihood: $\mathcal{L}(W, \mathcal{S}) = -\log \prod_{i=1}^{p} P(Y^i|X^i, W)$

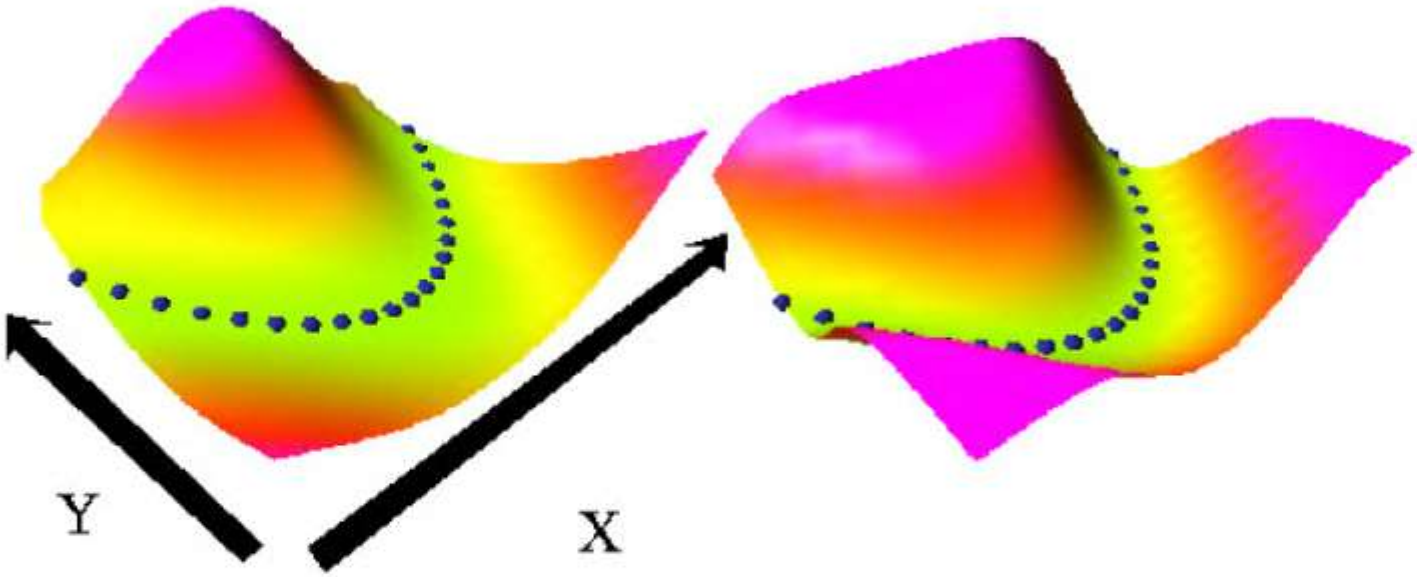
$$\mathcal{L}(W, \mathcal{S}) = \sum_{i=1}^{p} -\log\left(P(W, Y^i, X^i)\right) + \log\left(\int_y P(W, y, X^i)\right)$$

$$\mathcal{L}(W, \mathcal{S}) = \sum_{i=1}^{p} -\log\left(\int_z P(W, Y^i, z, X^i)\right) + \log\left(\int_{yz} P(W, y, z, X^i)\right)$$

What's so bad about probabilistic models?

- Why bother with a normalization since we don't use it for decision making?
- Why insist that P(Y|X) have a specific shape, when we only care about the position of its minimum?
- When Y is high-dimensional (or simply conbinatorial), normalizing becomes intractable (e.g. Language modeling, image restoration, large DoF robot control...).
- A tiny number of models are pre-normalized (Gaussian, exponential family)
- A very small number are easily normalizable
- A large number have intractable normalization
- A huuuge number can't be normalized at all (examples will be shown).
- Normalization forces us to take into account areas of the space that we don't actually care about because our inference algorithm never takes us there.
- **If we only care about making the right decisions, maximizing the likelihood solves a much more complex problem than we have to.**

EBM Energy Surfaces

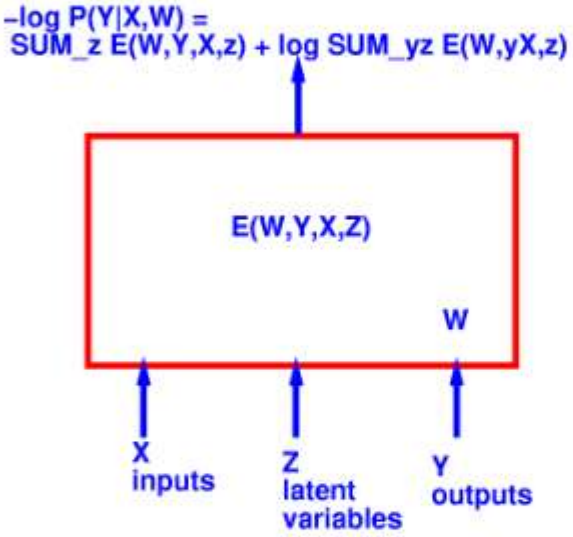


Examples: An EBM that computes $Y = X^2$.

On the left: $E(Y, X)$ is quadratic in Y. It corresponds to a Gaussian model of $P(Y|X)$.

On the right: $E(Y, X)$ is saturated. Although it gives the same answers as the EBM on the left, it has no probabilistic equivalent because the partition function $\int_y \exp(-E(Y, X))$ does not converge.

Probabilistic Models from Energy-Based Models



- Any joint probability model can be approached as close as we want by an equivalent EBM. If $P(Y, X, Z|W)$ is non-zero everywhere:
  $E(W, Y, X, Z) = C - \frac{1}{\beta} \log P(Y, X, Z|W)$
  where C is an arbitrary constant and β a strictly positive constant.
- not all EBMs can be turned into a probabilistic model. Only those for which $\int_{yz} \exp(-\beta E(W, y, X, z))$ converges:

$$P(Y|X) = \frac{\int_z \exp(-\beta E(W, Y, X, z))}{\int_{yz} \exp(-\beta E(W, y, X, z))}$$

Any single probabilistic model will have many equivalent EBMs when it comes to comparison-based inference or decision. *Because many energy surfaces have minima at the same places.*

We have a lot more flexibility with EBMs than with Prob. Models

What's good/bad about EBM?

What's bad about EBMs:

- There is no compositionality...
- ... but we don't care because we are going to train our whole system ***end-to-end***. With *end-to-end learning*, we do not need compositionality.

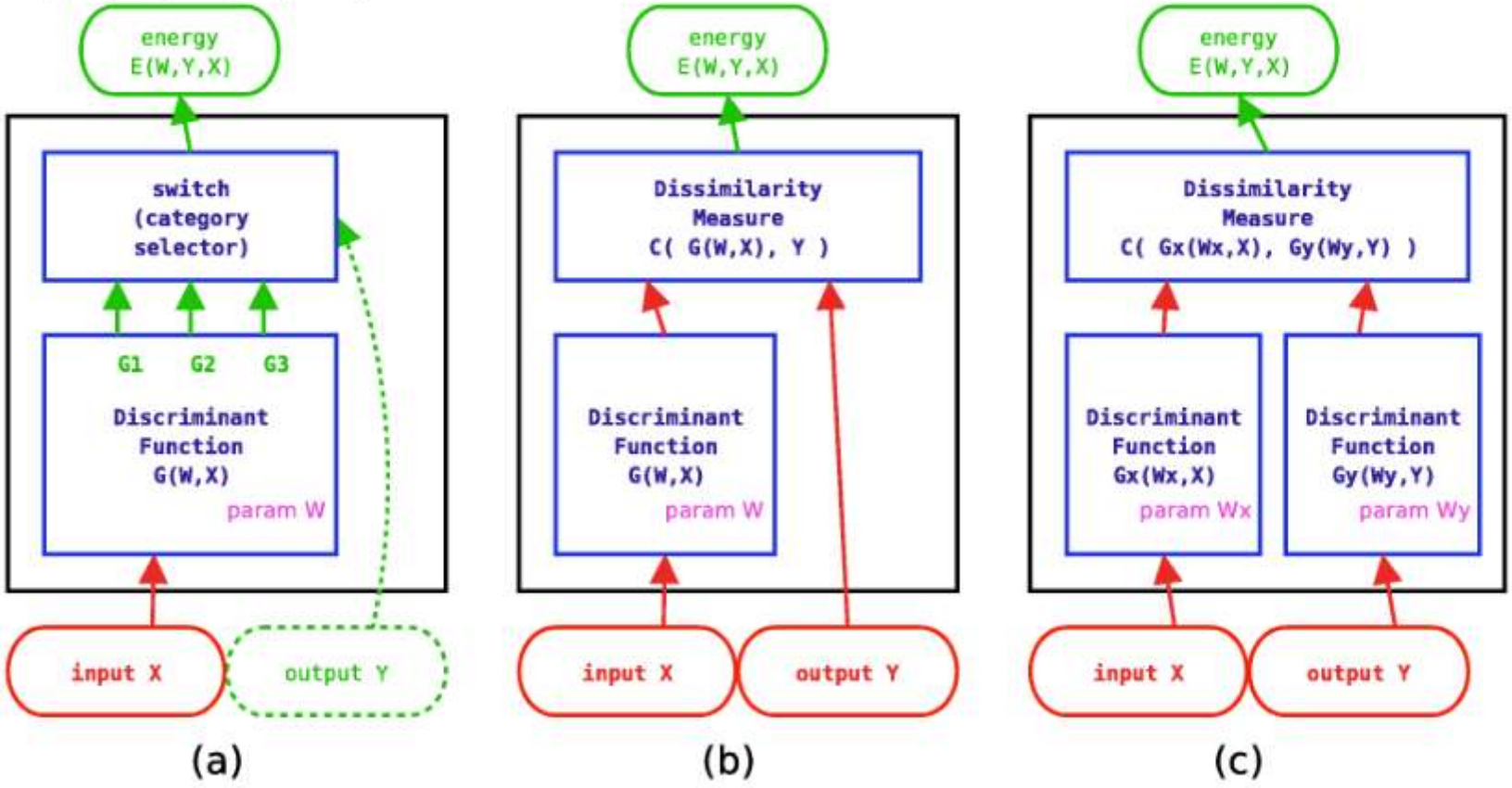
What's good about EBMs:

- We have complete freedom for the form and parameterization of the energy function (including things that can't be normalized).
- because we do not need to normalize, we can use a much larger repertoire of model architectures.
- No need for computing (intractable) partition functions
- No need to justify the choice of your favorite approximation of the partition function.

Pretty much every model we know is some form of EBM.
QUESTION: **what loss functions can we use for training?**.

Examples of EBM

Almost every type of model we know is some form of EBM. It all depends on how $E(W, Y, X, Z)$ is parameterized. NOTE: there is no linearity assumption.



Training EBMs

- Training will consist in finding a W that minimizes a loss function $\mathcal{L}(W, \mathcal{S})$, over the training set $\mathcal{S}$.
- We must devise loss functions that **"carve"** the energy landscape so that the energy is small around training samples and high everywhere else..
- We seek loss functions that **do not require** evaluating intractable integrals, but which, nevertheless, drive the machine to approach the desired behavior.
- Basic idea: **"dig holes"** at (X, Y) locations near training samples, while **"building hills"** at un-desired locations, particularly the ones that are erroneously picked by the inference algorithm.
- Whereas probabilistic models trained with max likelihood shape the entire energy surface, our EBM loss function will merely dig holes at the right places and build hills only where needed to avoid erroneous inferences.

Loss Functions for EBMs

- training set $\mathcal{S} = \{(X^i, Y^i)\ ,\ i = 1..p\}$
- Loss:

$$\mathcal{L}(W, \mathcal{S}) = R\left(\frac{1}{p}\sum_{i=1}^{p} L(W, Y^i, X^i)\right)$$

- $L(W, Y^i, X^i)$ is the per-sample loss function for sample (X^i, Y^i). L is assumed to have a lower bound.
- R is a monotonically increasing function. In the following we assume R=identity
- the loss is invariant under permutations of the samples, and under multiple repetitions of the same training set.
- What form can $L(W, Y, X)$ take?

Conditions on the Loss

- Condition for correct output on sample (X^i, Y^i): there is a margin $m > 0$, such that:

$$\check{E}(W, Y^i, X^i) < \check{E}(W, Y, X^i) - m \ , \quad \forall Y \in \{Y\}, Y \neq Y^i$$

- Assumption: L depends on X^i only through the set of energies $\{\check{E}(W, Y, X^i) \ , \ Y \in \{Y\}\}$.
- For example, if $\{Y\} = \{0, 1, \ldots, k-1\}$

$$L(W, Y^i, X^i) = L(Y^i, \check{E}(W, 0, X^i), \ldots, \check{E}(W, k-1, X^i))$$

- We want to design L so that making an update of W to decrease $L(W, Y^i, X^i)$ will automatically decrease the difference $\check{E}(W, Y^i, X^i) - \check{E}(W, Y, X^i)$ for values of Y such that $\check{E}(W, Y^i, X^i) < \check{E}(W, Y, X^i) - m$.

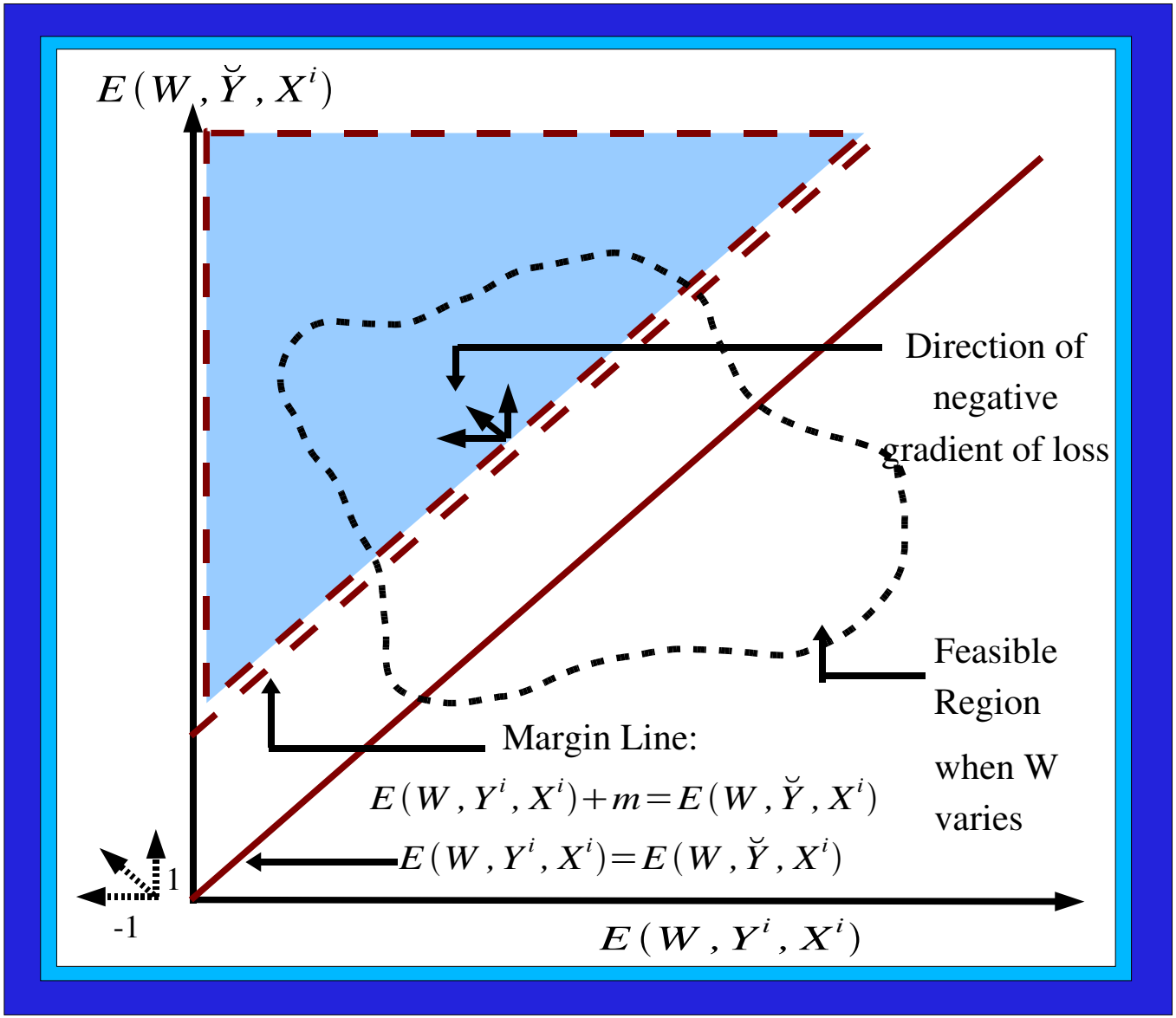
Conditions on the Loss

- Let's define $\bar{Y}$ as the most offending incorrect output:

$$\bar{Y} = \text{argmin}_{y \in \{Y\}, y \neq Y^i} \check{E}(W, y, X^i)$$

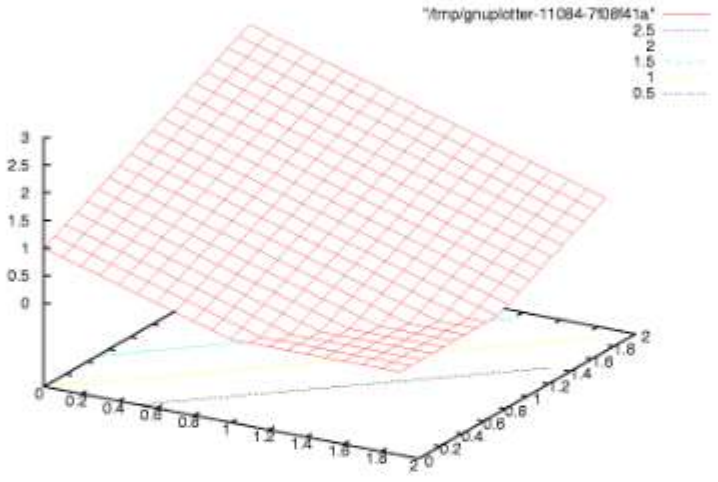
- Cond. for correct output: $\check{E}(W, Y^i, X^i) < \check{E}(W, \bar{Y}, X^i) - m$
- Let's assume that $L(W, Y^i, X^i)$ is convex in the 2 coordinates $\check{E}(W, Y^i, X^i)$ and $\check{E}(W, \bar{Y}, X^i)$.
- **Sufficient condition 1**: All the minima of $L(W, Y^i, X^i)$ must be in the half-plane $\check{E}(W, Y^i, X^i) < \check{E}(W, \bar{Y}, X^i) - m$.
- **Sufficient condition 2**: The gradient of $L(W, Y^i, X^i)$ on the margin line $\check{E}(W, Y^i, X^i) = \check{E}(W, \bar{Y}, X^i) - m$, must have a positive dot product with the direction [-1,1].
- **Sufficient condition 3**: On the margin line $\check{E}(W, \bar{Y}, X^i) = \check{E}(W, Y^i, X^i) + m$, the following must hold: $\left[\frac{\partial \check{E}(W,Y^i,X^i)}{\partial W} - \frac{\partial \check{E}(W,\bar{Y},X^i)}{\partial W}\right] \cdot \frac{\partial L(W,Y^i,X^i)}{\partial W} > 0$

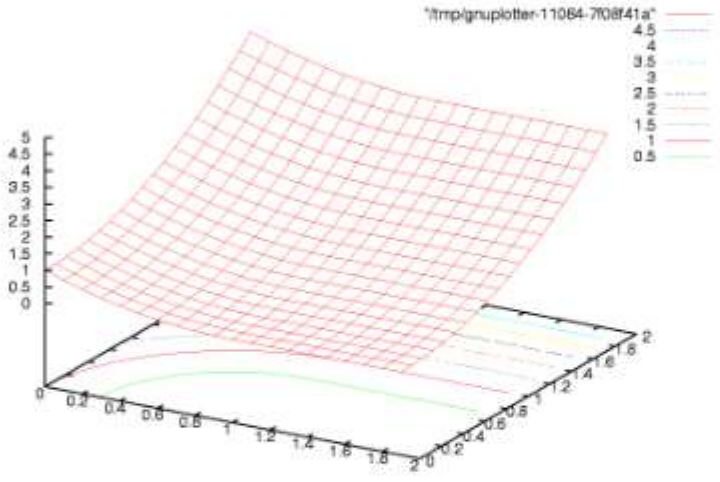
Conditions on the Loss

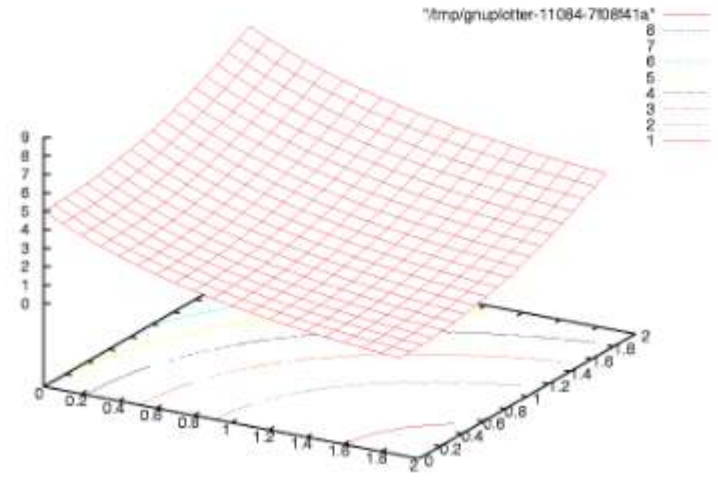


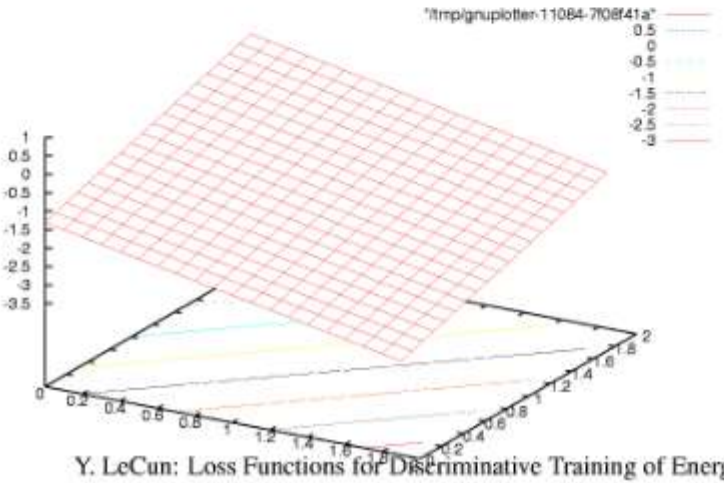
Condition on the Loss

Loss $L(W, Y^i, X^i)$ as a function of $\check{E}(W, \bar{Y}, X^i)$ and $\check{E}(W, Y^i, X^i)$







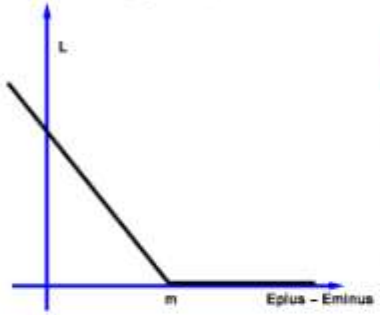


Examples of Loss Functions

- Energy Loss: $L_{\text{energy}}(W, Y^i, X^i) = \check{E}(W, Y^i, X^i)$.
  Only works if the architecture is such that decreasing $\check{E}(W, Y^i, X^i)$ will automatically increase $\check{E}(W, Y, X^i)$ for $y \neq Y^i$.
- Generalized Perceptron Loss [LeCun 1998][Collins 2002]:
  $L_{\text{ptron}}(W, Y^i, X^i) = \check{E}(W, Y^i, X^i) - \min_{Y \in \{Y\}} \check{E}(W, Y, X^i)$
  Does not work because the margin is zero. This reduces to the traditional linear perceptron loss when $\check{E}(W, Y, X) = -YW.X$.
- Generalized Margin Loss: [LeCun et al. 2005]
  $L_{\text{gmargin}}(W, Y^i, X^i) = Q[\check{E}(W, Y^i, X^i), \check{E}(W, \bar{Y}, X^i)]$
  Where Q is an increasing function of $\check{E}(W, Y^i, X^i)$ and a decreasing function of $\check{E}(W, \bar{Y}, X^i)$.
- Negative Log Likelihood Loss: [Bengio 1992][LeCun 1998][Lafferty 2001]
  $L_{\text{nll}}(W, Y^i, X^i) = \check{E}(W, Y^i, X^i) - F_\beta(W, X^i)$
  with: $F_\beta(W, X^i) = -\frac{1}{\beta} \log \left(\int_{Y \in \{Y\}} \exp[-\beta \check{E}(W, Y, X^i)] \right)$

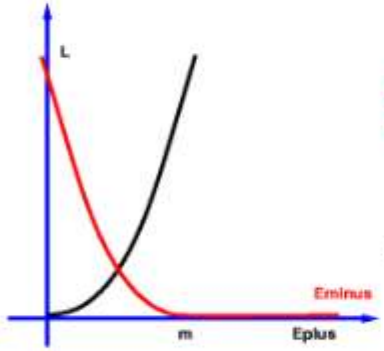
Special Cases of the Generalized Margin Loss

$$L_{\text{gmargin}}(W, Y^i, X^i) = Q[\check{E}(W, Y^i, X^i), \check{E}(W, \bar{Y}, X^i)]$$



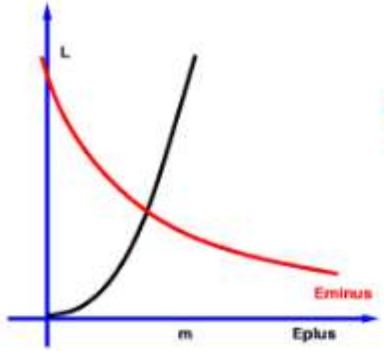
Hinge Loss [Taskar, Guestrin, Koller 2003],[Altun, Johnson, Hofmann, 2003]:

$$L_{\text{hinge}}(W, Y^i, X^i) = \max(0, m + \check{E}(W, Y^i, X^i) - \check{E}(W, \bar{Y}, X^i))$$



Square-Square Loss:

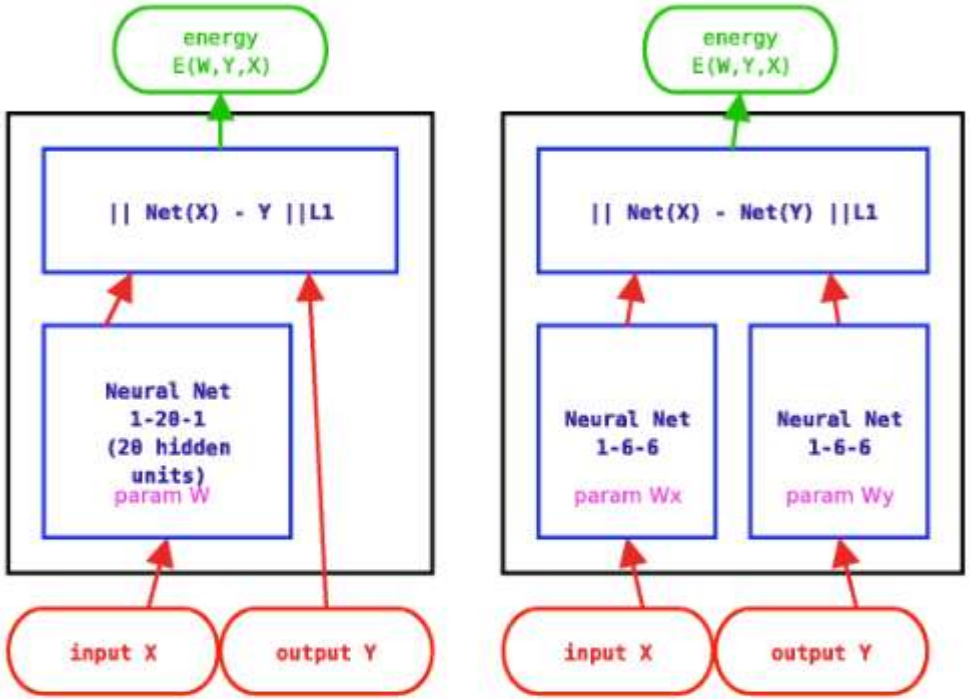
$$L_{\text{sqsq}}(W, Y^i, X^i) = \check{E}(W, Y^i, X^i)^2 + (\min(0, m - \check{E}(W, \bar{Y}, X^i)))^2$$



Square-Exp Loss: [Osadchy, Miller, LeCun, NIPS 2004]

$$L_{\text{sqexp}}(W, Y^i, X^i) = \check{E}(W, Y^i, X^i)^2 + K \exp(-\beta \check{E}(W, \bar{Y}, X^i))$$

EBM Demos

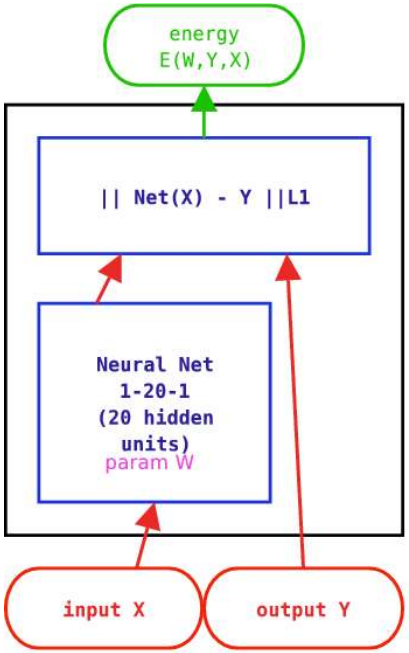


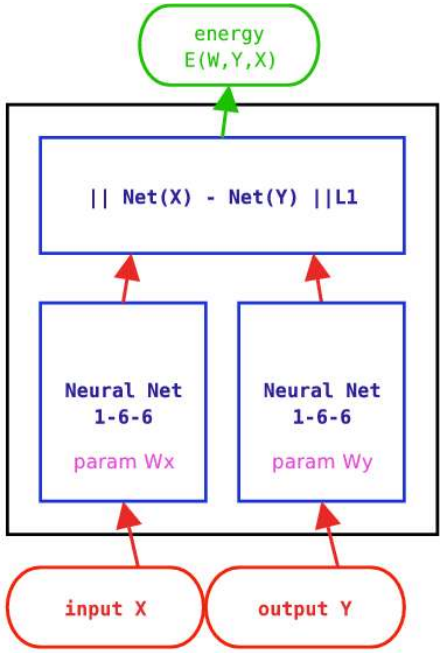
- Demo 1: $Y = X^2$, Architecture A, Square Energy Loss. It works because $E(Y, X)$ is a fixed quadratic function of Y.
- Demo 2: $Y = X^2$, Architecture B, Square Energy. It collapses.
- Demo 3: $Y = X^2$, Architecture B, Square-Square Margin Loss
- Demo 4: $Y = X^2$, Architecture B, Negative Log Likelihood Loss. Few iterations, but each iteration is expensive

Initially, the forbidden sphere around Y^i is 0.2, then 0.1.
Demo 5: eye pattern, Architecture B, Negative Log Likelihood Loss.

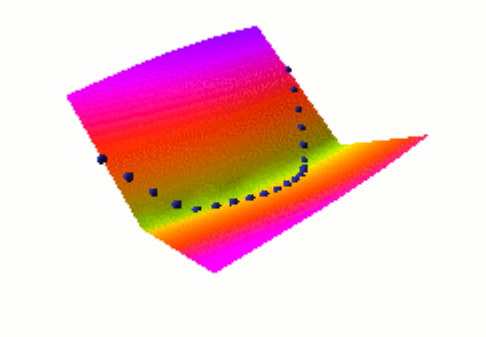
EBM Demos: energy loss

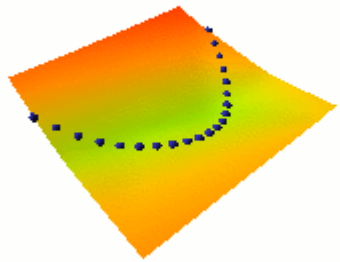
Loss: "Energy Loss": L(W,Y,X) = E(W,Y,X)





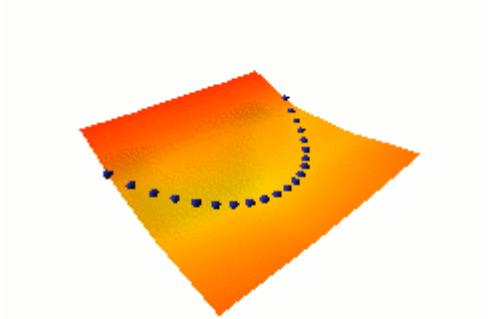
COLLAPSE!!!

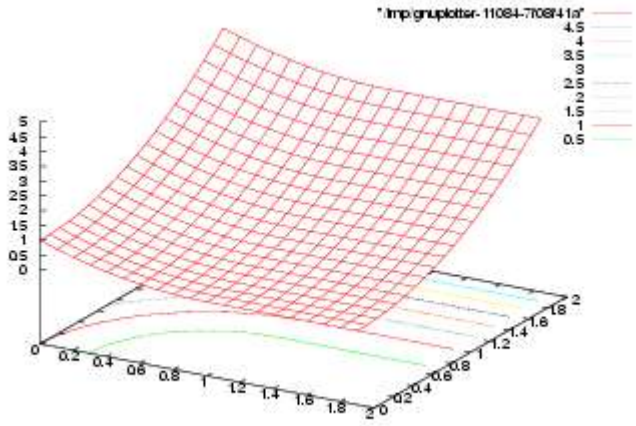


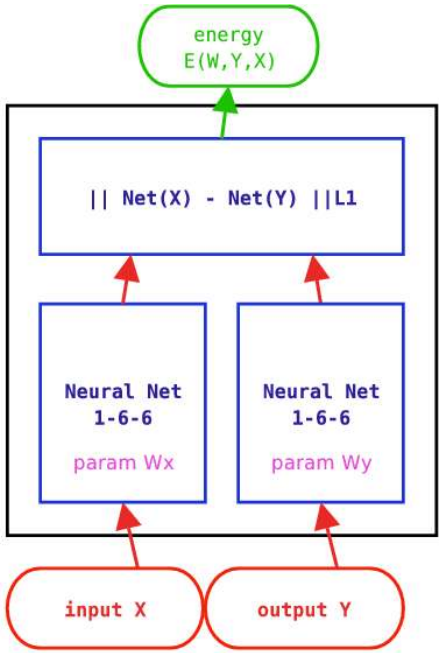


EBM Demos: good losses

Square-square loss

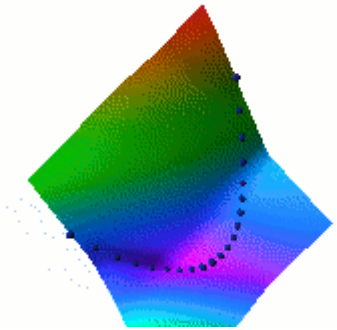


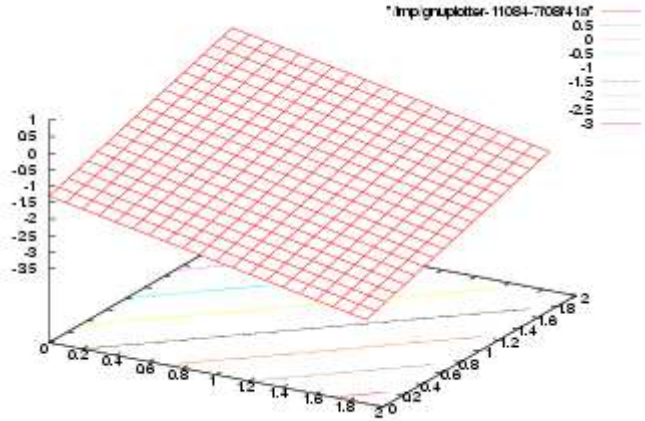




(b)

Neg-Log-Likelihood loss



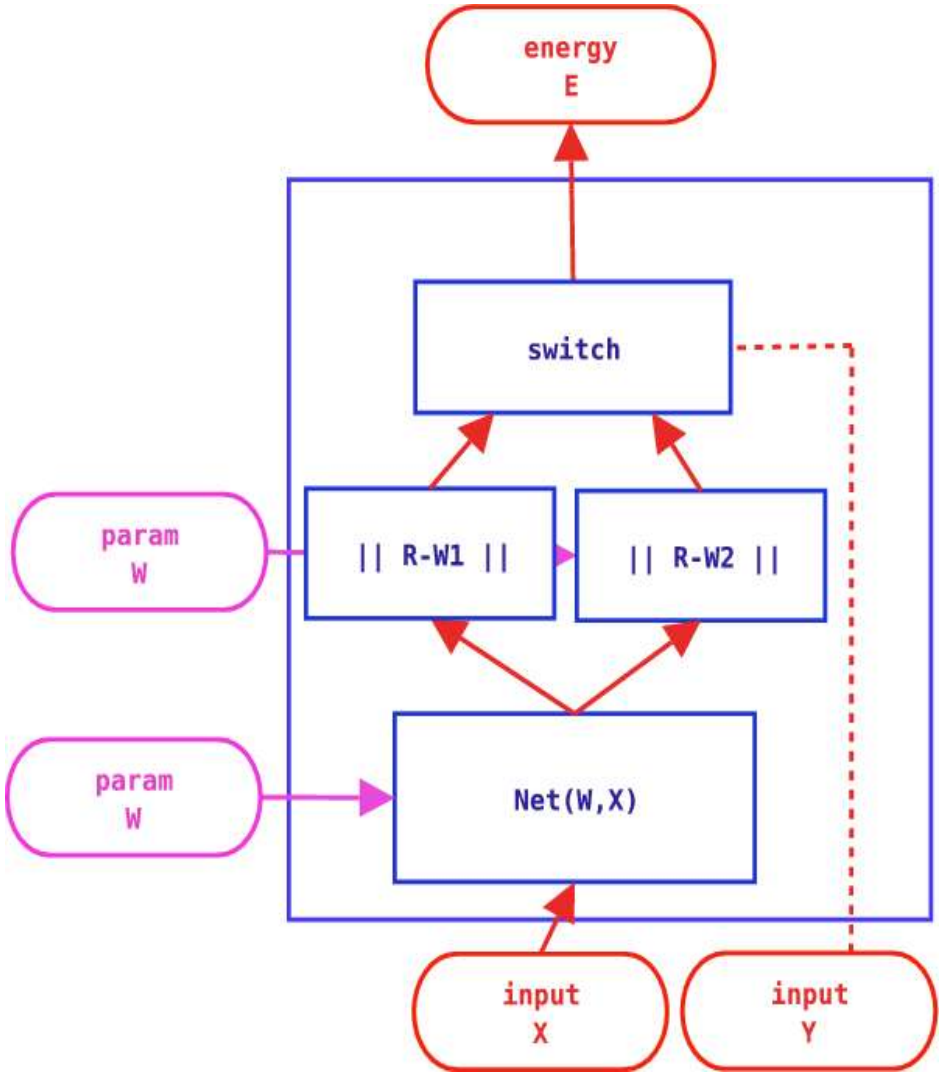


$$L(W,X,Y) = E(W,Y,X)^2 - \max\left(0, m - E(W,\bar{Y},X)^2\right)$$

$$\bar{Y} = \text{argmin}_{y\in\{Y\},y\neq Y} E(W,y,X)$$

$$L(W,X,Y) = E(W,Y,X) + \frac{1}{\beta}\log\left[\sum_{y\in\{Y\}}\exp\left(-\beta E(W,y,X)\right)\right]$$

Other Architectures that may collapse



- **Linear module followed by radial basis functions and a switch:**
  - Will collapse with the energy loss and the perceptron loss.
  - Will not collapse with the square-square, neg-log-likelihood, margin loss, etc....

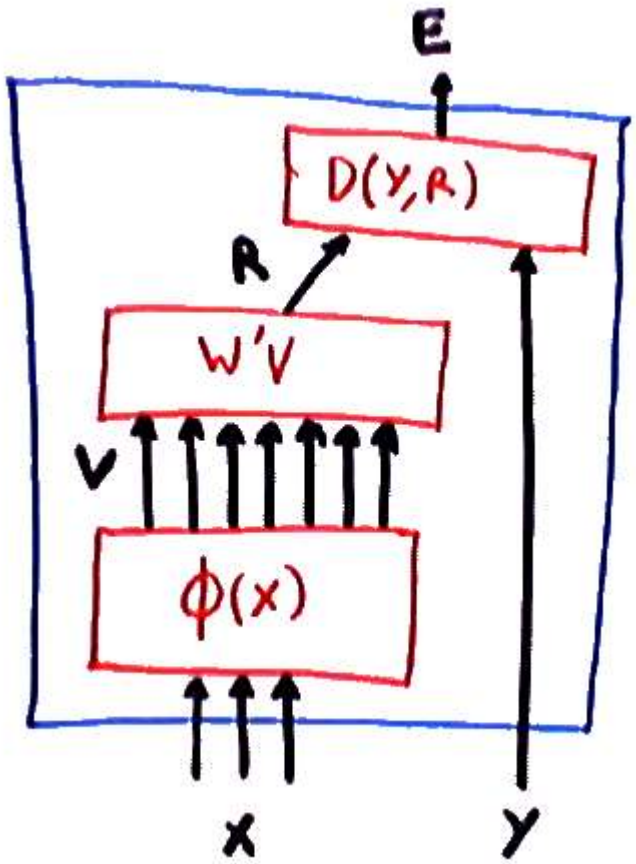
EBM

- Unlike traditional classifiers, EBMs can represent multiple alternative outputs
- The normalization in probabilistic models is often an unnecessary aggravation, particularly if the ultimate goal of the system is to make decisions.
- EBMs with appropriate loss function avoid the necessity to compute the partition function and its derivatives (which may be intractable)
- EBMs give us complete freedom in the choice of the architecture that models the joint "incompatibility" (energy) between the variables.
- We can use architectures that are not normally allowed in the probabilistic framework (like neural nets).
- **The inference algorithm that finds the most offending (lowest energy) incorrect answer does not need to be exact:** our model may give low energy to far-away regions of the landscape. But if our inference algorithm never finds those regions, **they do not affect us.** But they do affect normalized probabilistic models

Non-Linear Models

- **We can always add "features", "kernels", or "basis function" in front of a linear model to make it non-linear.**
  - This is how non-linear SVMs are built.
- **Question: so, why would we need anything else?**
- **Answer: the complexity of the real world is very difficult to capture in a kernel.**
- **How do we solve:**
  - The invariance problem in image recognition.
  - The structured output problem in sequence labeling (parts of speech tagging, speech recognition, biological sequence analysis....).

Fixed Preprocessing (features, kernels, basis functions)

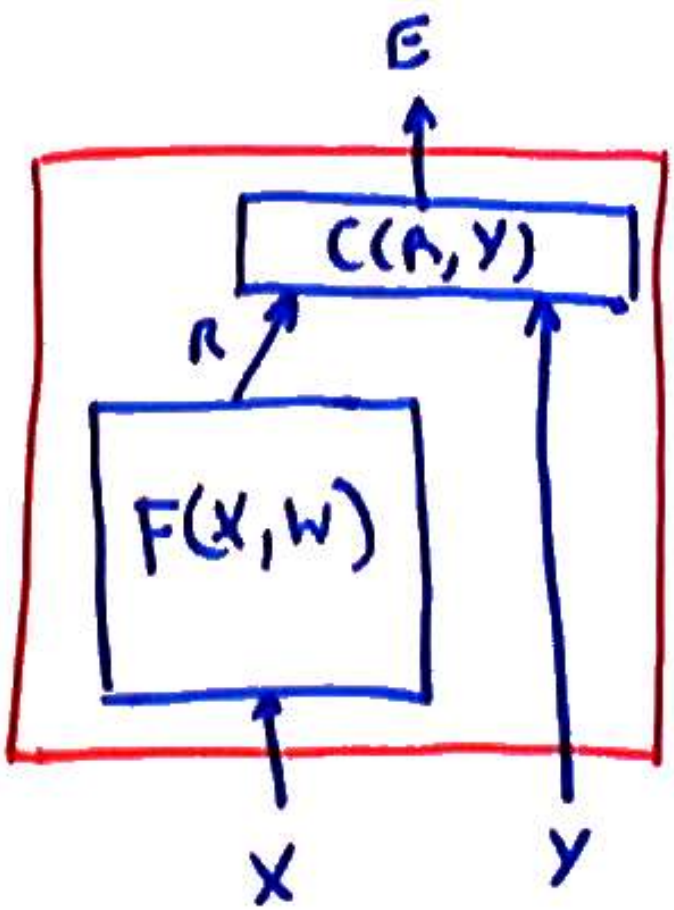


- **Simplest approach:**
  - Make each basis function a Gaussian bump (a template matcher).
  - Put one bump centered on each training sample.
  - In the space on the which the linear parameter operate, we get one separate dimension for each training sample (we can learn anything).
  - Regularize.
  - You get a **Support Vector Machine.**
- **Problem: an SVM is a glorified template matcher which is only as good as its kernel.**

Trainable Front-End, Structured Architectures

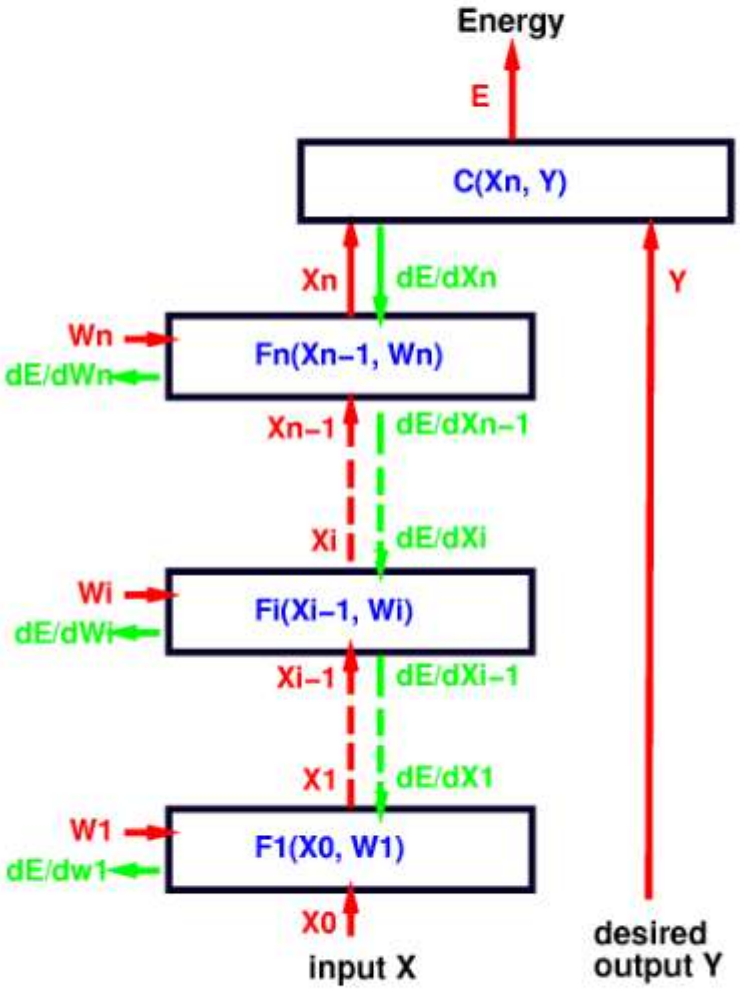
- **The Solutions:**
  - Invariance: Do not use a fixed front-end, **make it trainable**, so it can learn to extract invariant representations
  - Structure: Do not use a simple linearly-parameterized classifier, use architectures whose inference process involves search and "reasoning" (a Hidden Markov Model, a Markov Random Field, a general graphical model).
- **We need total flexibility in the design of the architecture of the machine:**
  - So that we can tailor the architecture to the task
  - So that we can build our prior knowledge about the task into the architecture
- **Multi-Module Architectures.**

Multi-Module Architectures



- **For Supervised Learning**
  - We allow the function F(W,X) to be non-linearly parameterized in W.
  - This allows us to play with a large repertoire of functions with rich class boundaries.
  - We assume that F(W,X) is differentiable almost everywhere with respect to W.

Multi-Module Systems: Cascade

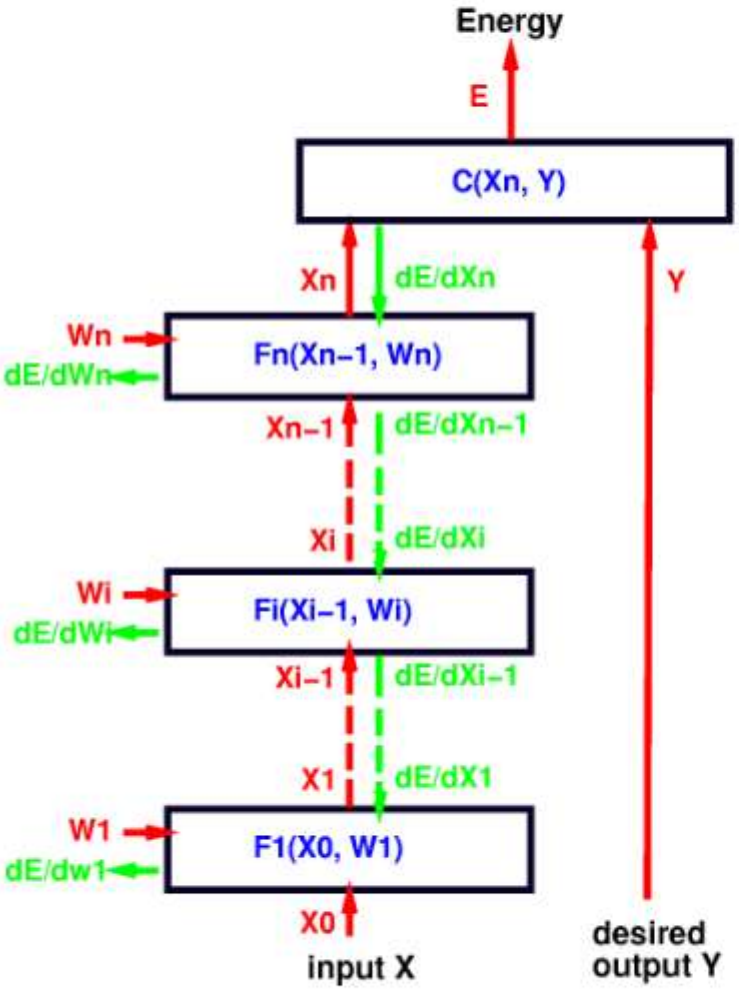


- Complex learning machines can be built by assembling Modules into networks.
- a simple example: layered, feed-forward architecture (cascade).
- computing the output from the input: forward propagation
- let $X = X_0$,

$$X_i = F_i(X_{i-1}, W_i) \quad \forall i \in [1, n]$$

$$E(Y, X, W) = C(X_n, Y)$$

Object-Oriented Implementation



- Each module is an object (instance of a class).
- Each class has an "fprop" (forward propagation) method that takes the input and output states as arguments and computes the output state from the input state.
- Lush:
  ```
  (==> module fprop input output)
  ```
- C++:
  ```
  module.fprop(input,output);
  ```

Gradient of the Loss, gradient of the Energy

- We assumed early on that the loss depends on W only through the terms $E(W, Y, X^i)$:

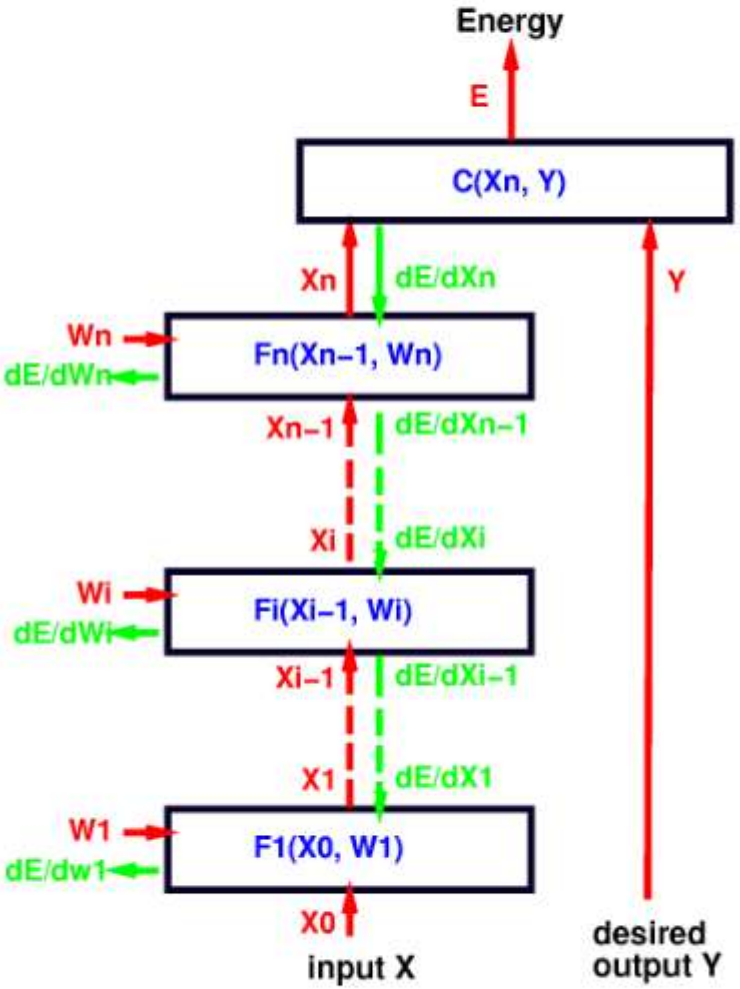
$$L(W, Y^i, X^i) = L(Y^i, E(W, 0, X^i), E(W, 1, X^i), \ldots, E(W, k-1, X^i))$$

- therefore:

$$\frac{\partial L(W, Y^i, X^i)}{\partial W} = \sum_Y \frac{\partial L(W, Y^i, X^i)}{\partial E(W, Y, X^i)} \frac{\partial E(W, Y, X^i)}{\partial W}]$$

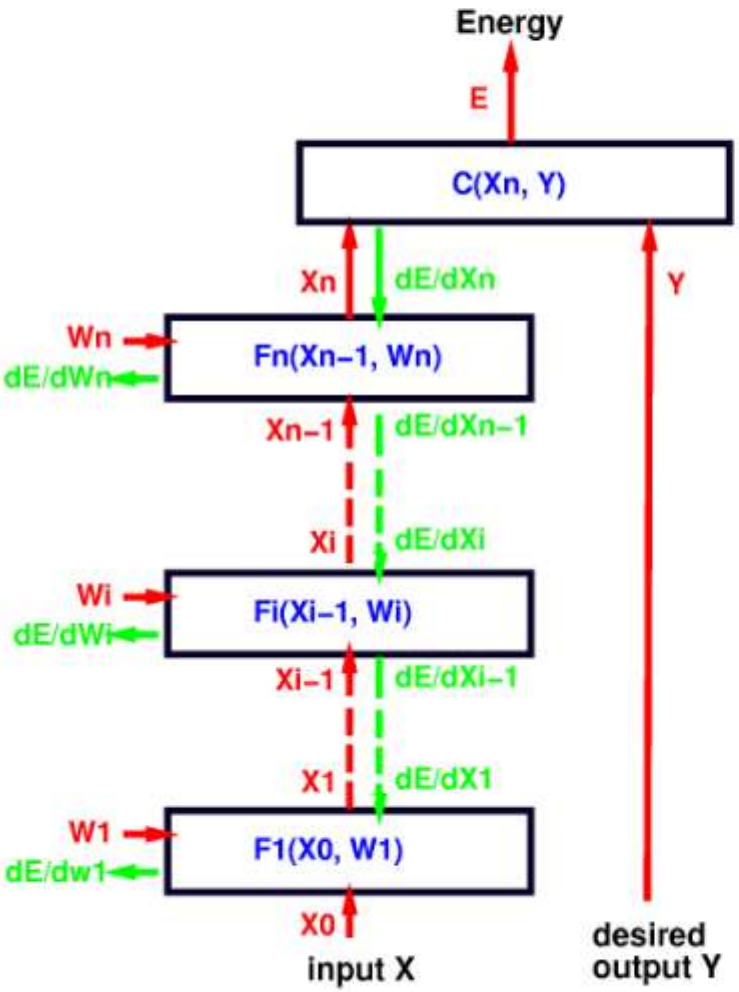
- We only need to compute the terms $\frac{\partial E(W,Y,X^i)}{\partial W}$
- Question: How do we compute those terms efficiently?

Computing the Gradients in Multi-Layer Systems



- To train a multi-module system, we must compute the gradient of E with respect to all the parameters in the system (all the W_i).
- Let's consider module i whose fprop method computes $X_i = F_i(X_{i-1}, W_i)$.
- Let's assume that we already know $\frac{\partial E}{\partial X_i}$, in other words, for each component of vector X_i we know how much E would wiggle if we wiggled that component of X_i.

Computing the Gradients in Multi-Layer Systems



- We can apply chain rule to compute $\frac{\partial E}{\partial W_i}$ (how much E would wiggle if we wiggled each component of W_i):

$$\frac{\partial E}{\partial W_i} = \frac{\partial E}{\partial X_i} \frac{\partial F_i(X_{i-1}, W_i)}{\partial W_i}$$

$$[1 \times N_w] = [1 \times N_x].[N_x \times N_w]$$

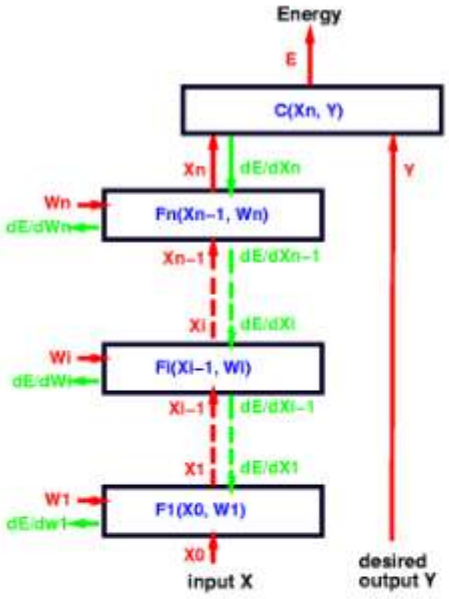
- $\frac{\partial F_i(X_{i-1}, W_i)}{\partial W_i}$ is the *Jacobian matrix* of F_i with respect to W_i.

$$\left[\frac{\partial F_i(X_{i-1}, W_i)}{\partial W_i}\right]_{kl} = \frac{\partial \left[F_i(X_{i-1}, W_i)\right]_k}{\partial [W_i]_l}$$

- Element (k, l) of the Jacobian indicates how much the k-th output wiggles when we wiggle the l-th weight.

Computing the Gradients in Multi-Layer Systems

Using the same trick, we can compute $\frac{\partial E}{\partial X_{i-1}}$. Let's assume again that we already know $\frac{\partial E}{\partial X_i}$, in other words, for each component of vector X_i we know how much E would wiggle if we wiggled that component of X_i.



- We can apply chain rule to compute $\frac{\partial E}{\partial X_{i-1}}$ (how much E would wiggle if we wiggled each component of X_{i-1}):

$$\frac{\partial E}{\partial X_{i-1}} = \frac{\partial E}{\partial X_i} \frac{\partial F_i(X_{i-1}, W_i)}{\partial X_{i-1}}$$

- $\frac{\partial F_i(X_{i-1}, W_i)}{\partial X_{i-1}}$ is the *Jacobian matrix* of F_i with respect to X_{i-1}.
- F_i has two Jacobian matrices, because it has to arguments.
- Element (k, l) of this Jacobian indicates how much the k-th output wiggles when we wiggle the l-th input.
- **The equation above is a recurrence equation!**

Jacobians and Dimensions

- derivatives with respect to a column vector are line vectors (dimensions: $[1 \times N_{i-1}] = [1 \times N_i] * [N_i \times N_{i-1}]$)

$$\frac{\partial E}{\partial X_{i-1}} = \frac{\partial E}{\partial X_i} \frac{\partial F_i(X_{i-1}, W_i)}{\partial X_{i-1}}$$

- (dimensions: $[1 \times N_{wi}] = [1 \times N_i] * [N_i \times N_{wi}]$):

$$\frac{\partial E}{\partial W_i} = \frac{\partial E}{\partial X_i} \frac{\partial F_i(X_{i-1}, W_i)}{\partial W}$$

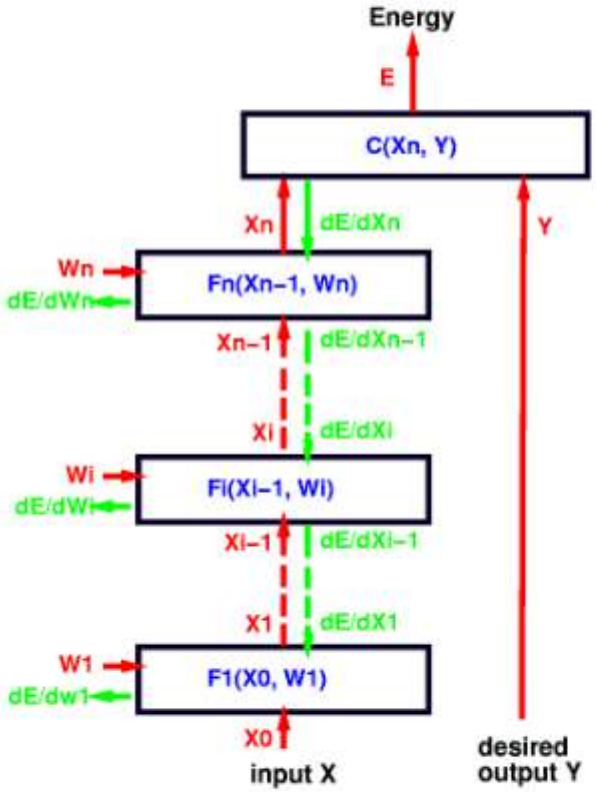
- we may prefer to write those equation with column vectors:

$$\frac{\partial E}{\partial X_{i-1}}' = \frac{\partial F_i(X_{i-1}, W_i)}{\partial X_{i-1}}' \frac{\partial E}{\partial X_i}'$$

$$\frac{\partial E}{\partial W_i}' = \frac{\partial F_i(X_{i-1}, W_i)}{\partial W}' \frac{\partial E}{\partial X_i}'$$

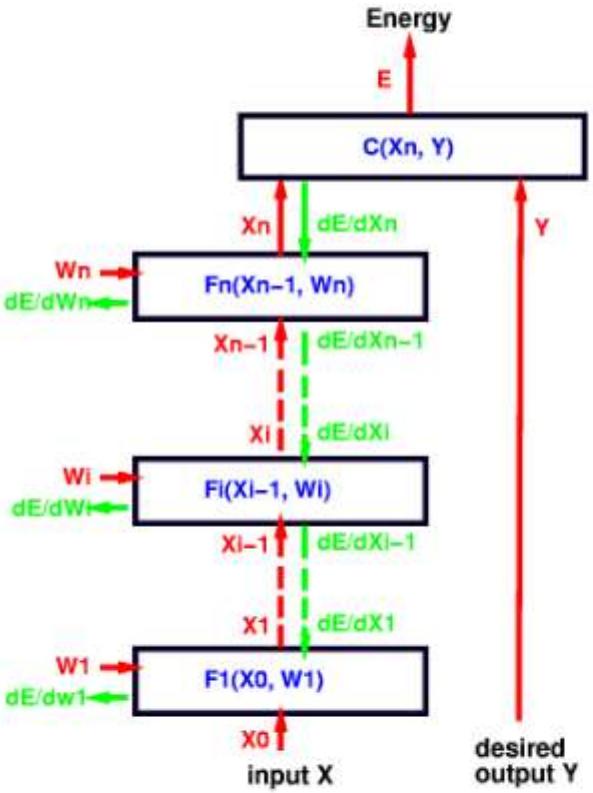
Back-propagation

To compute all the derivatives, we use a backward sweep called the **back-propagation algorithm** that uses the recurrence equation for $\frac{\partial E}{\partial X_i}$



- $\frac{\partial E}{\partial X_n} = \frac{\partial C(X_n, Y)}{\partial X_n}$
- $\frac{\partial E}{\partial X_{n-1}} = \frac{\partial E}{\partial X_n} \frac{\partial F_n(X_{n-1}, W_n)}{\partial X_{n-1}}$
- $\frac{\partial E}{\partial W_n} = \frac{\partial E}{\partial X_n} \frac{\partial F_n(X_{n-1}, W_n)}{\partial W_n}$
- $\frac{\partial E}{\partial X_{n-2}} = \frac{\partial E}{\partial X_{n-1}} \frac{\partial F_{n-1}(X_{n-2}, W_{n-1})}{\partial X_{n-2}}$
- $\frac{\partial E}{\partial W_{n-1}} = \frac{\partial E}{\partial X_{n-1}} \frac{\partial F_{n-1}(X_{n-2}, W_{n-1})}{\partial W_{n-1}}$
- ....etc, until we reach the first module.
- we now have all the $\frac{\partial E}{\partial W_i}$ for $i \in [1, n]$.

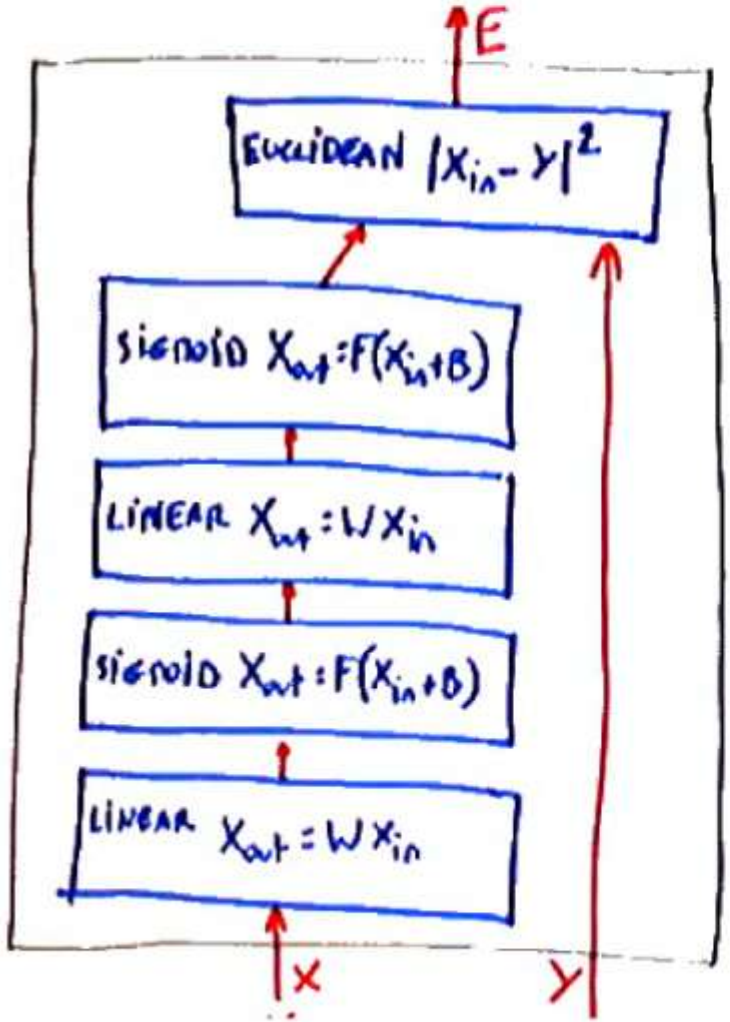
Object-Oriented Implementation



- Each module is an object (instance of a class).
- Each class has a "bprop" (backward propagation) method that takes the input and output states as arguments and computes the derivative of the energy with respect to the input from the derivative with respect to the output:
- Lush: `(==> module bprop input output)`
- C++: `module.bprop(input,output);`
- the objects `input` and `output` contain two slots: one vector for the forward state, and one vector for the backward derivatives.
- the method `bprop` computes the backward derivative slot of `input`, by multiplying the backward derivative slot of `output` by the Jacobian of the module at the forward state of `input`.

Modules in a Multi-layer Neural Net

A fully-connected, feed-forward, multi-layer neural nets can be implemented by stacking three types of modules.

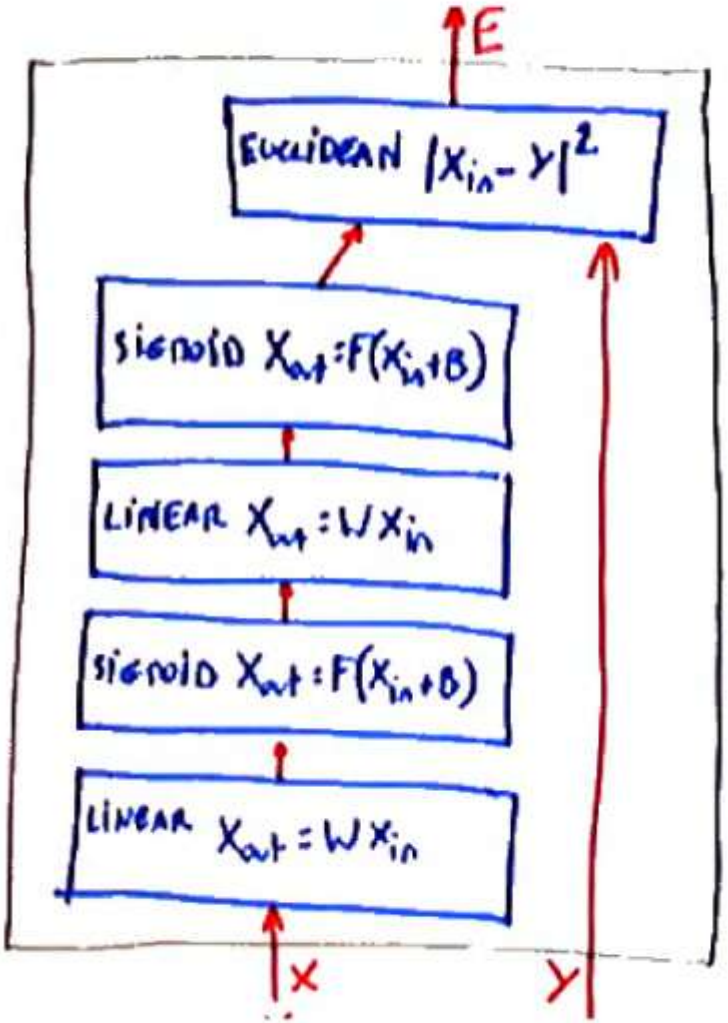


- Linear modules: X_{in} and X_{out} are vectors, and W is a weight matrix.

$$X_{\text{out}} = W X_{\text{in}}$$

- Sigmoid modules: $(X_{\text{out}})_i = \sigma((X_{\text{in}})_i + B_i)$ where B is a vector of trainable "biases", and σ is a sigmoid function such as tanh or the logistic function.
- a Euclidean Distance module $E = \frac{1}{2}||Y - X_{\text{in}}||^2$. With this energy function, we will use the neural network as a regressor rather than a classifier.

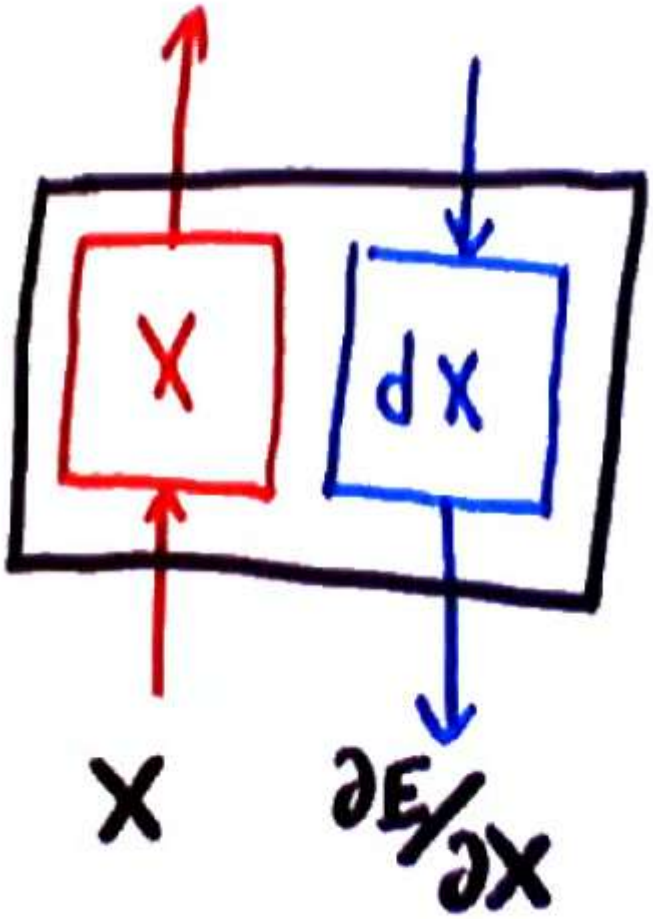
Loss Function



Here, we will us the simple Energy Loss function L_{energy}:

$$L_{\text{energy}}(W, Y^i, X^i) = E(W, Y^i, X^i)$$

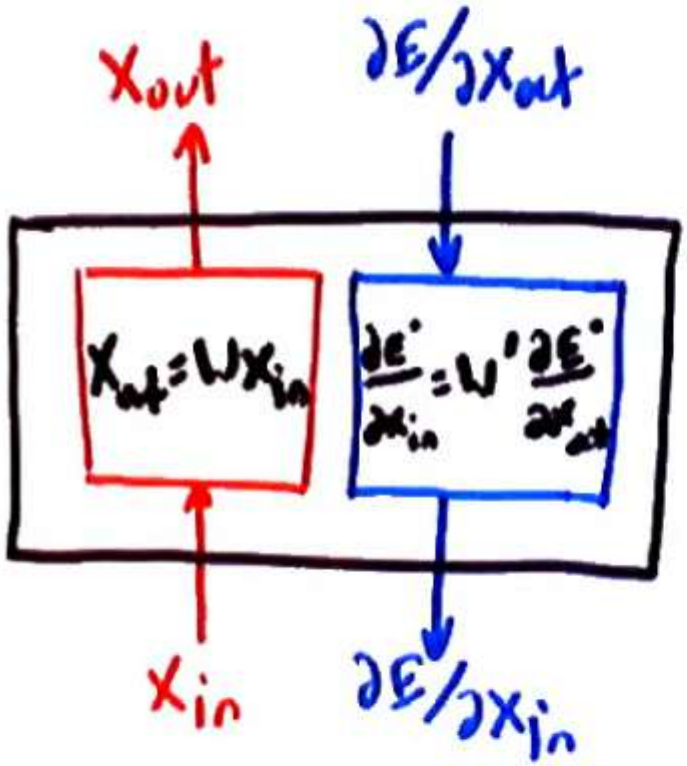
OO Implementation: the `state1` Class



the internal state of the network will be kept in a "state" class that contains two scalars, vectors, or matrices: (1) the state proper, (2) the derivative of the energy with respect to that state.

Linear Module

The input vector is multiplied by the weight matrix.



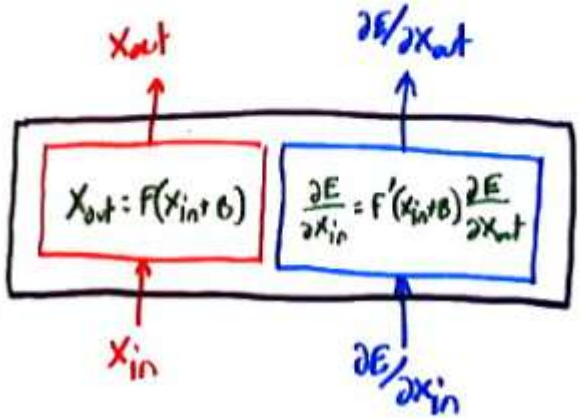
- fprop: $X_{\text{out}} = W X_{\text{in}}$
- bprop to input:
  $\frac{\partial E}{\partial X_{\text{in}}} = \frac{\partial E}{\partial X_{\text{out}}} \frac{\partial X_{\text{out}}}{\partial X_{\text{in}}} = \frac{\partial E}{\partial X_{\text{out}}} W$
- by transposing, we get column vectors:
  $\frac{\partial E}{\partial X_{\text{in}}}' = W' \frac{\partial E}{\partial X_{\text{out}}}'$
- bprop to weights:
  $\frac{\partial E}{\partial W_{ij}} = \frac{\partial E}{\partial X_{\text{out}i}} \frac{\partial X_{\text{out}i}}{\partial W_{ij}} = X_{\text{in}j} \frac{\partial E}{\partial X_{\text{out}i}}$
- We can write this as an outer-product:
  $\frac{\partial E}{\partial W}' = \frac{\partial E}{\partial X_{\text{out}}}' X_{in}'$

Linear Module

Lush implementation:

```
(defclass linear-module object w)
(defmethod linear-module linear-module (ninputs noutputs)
  (setq w (matrix noutputs ninputs)))
(defmethod linear-module fprop (input output)
  (==> output resize (idx-dim :w:x 0))
  (idx-m2dotm1 :w:x :input:x :output:x) ())
(defmethod linear-module bprop (input output)
  (idx-m2dotm1 (transpose :w:x) :output:dx :input:dx)
  (idx-m1extm1 :output:dx :input:x :w:dx) ())
```

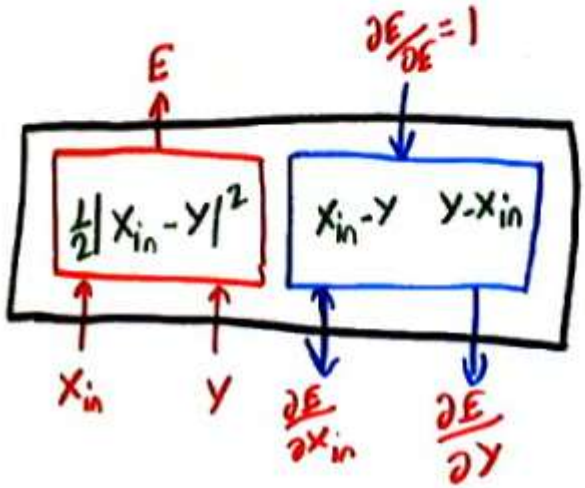
Sigmoid Module (tanh: hyperbolic tangent)



- fprop: $(X_{\text{out}})_i = \tanh((X_{\text{in}})_i + B_i)$
- bprop to input:
  $(\frac{\partial E}{\partial X_{\text{in}}})_i = (\frac{\partial E}{\partial X_{\text{out}}})_i \tanh'((X_{\text{in}})_i + B_i)$
- bprop to bias:
  $\frac{\partial E}{\partial B_i} = (\frac{\partial E}{\partial X_{\text{out}}})_i \tanh'((X_{\text{in}})_i + B_i)$
- $\tanh(x) = \frac{2}{1+\exp -x} - 1 = \frac{1-\exp(-x)}{1+\exp(-x)}$

```
(defclass tanh-module object bias)
(defmethod tanh-module tanh-module l
  (setq bias (apply matrix l)))
(defmethod tanh-module fprop (input output)
  (==> output resize (idx-dim :bias:x 0))
  (idx-add :input:x :bias:x :output:x)
  (idx-tanh :output:x :output:x))
(defmethod tanh-module bprop (input output)
  (idx-dtanh (idx-add :input:x :bias:x) :input:dx)
  (idx-mul :input:dx :output:dx :input:dx)
  (idx-copy :input:dx :bias:dx) ())
```

Euclidean Module



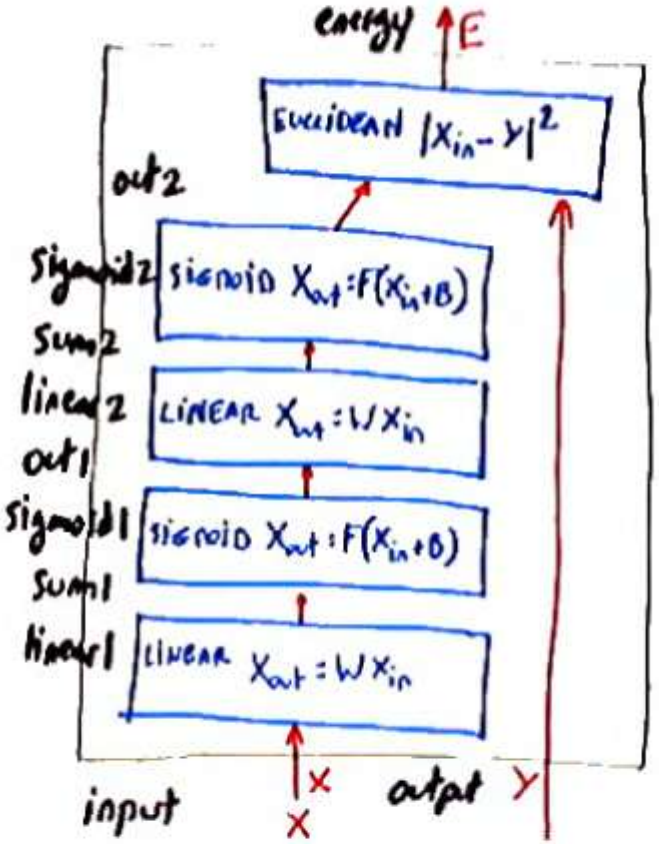
- fprop: $X_{\text{out}} = \frac{1}{2}||X_{\text{in}} - Y||^2$
- bprop to X input: $\frac{\partial E}{\partial X_{\text{in}}} = X_{\text{in}} - Y$
- bprop to Y input: $\frac{\partial E}{\partial Y} = Y - X_{\text{in}}$

```
(defclass euclidean-module object)
(defmethod euclidean-module run (input1 input2 output)
  (idx-copy :input1:x :input2:x)
  (:output:x 0) ())
(defmethod euclidean-module fprop (input1 input2 output)
  (idx-sqrdist :input1:x :input2:x :output:x)
  (:output:x (* 0.5 (:output:x))) ())
(defmethod euclidean-module bprop (input1 input2 output)
  (idx-sub :input1:x :input2:x :input1:dx)
  (idx-dotm0 :input1:dx :output:dx :input1:dx)
  (idx-minus :input1:dx :input2:dx))
```

Assembling the Network: A single layer

```
;; One layer of a neural net
(defclass nn-layer object
  linear  ; linear module
  sum  ; weighted sums
  sigmoid ; tanh-module
  )
(defmethod nn-layer nn-layer (ninputs noutputs)
  (setq linear (new linear-module ninputs noutputs))
  (setq sum (new state noutputs))
  (setq sigmoid (new tanh-module noutputs)) ())
(defmethod nn-layer fprop (input output)
  (==> linear fprop input sum)
  (==> sigmoid fprop sum output) ())
(defmethod nn-layer bprop (input output)
  (==> sigmoid bprop sum output)
  (==> linear bprop input sum) ())
```

Assembling a 2-layer Net



- Class implementation for a 2 layer, feed forward neural net.

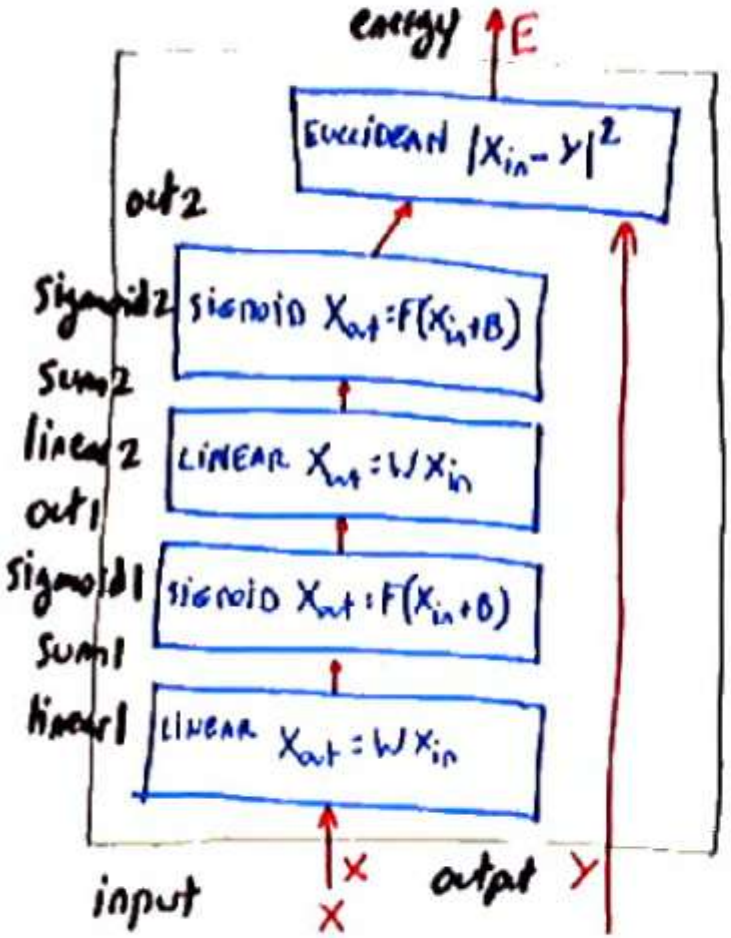
```
(defclass nn-2layer object
  layer1  ; first layer module
  hidden  ; hidden state
  layer2 ; second layer
  )
(defmethod nn-2layer nn-2layer (ninputs nhi
  (setq layer1 (new nn-layer ninputs nhidde
  (setq hidden (new state nhidden))
  (setq layer2 (new nn-layer nhidden noutpu
```

Assembling the Network: fprop and bprop

Implementation of a 2 layer, feed forward neural net.

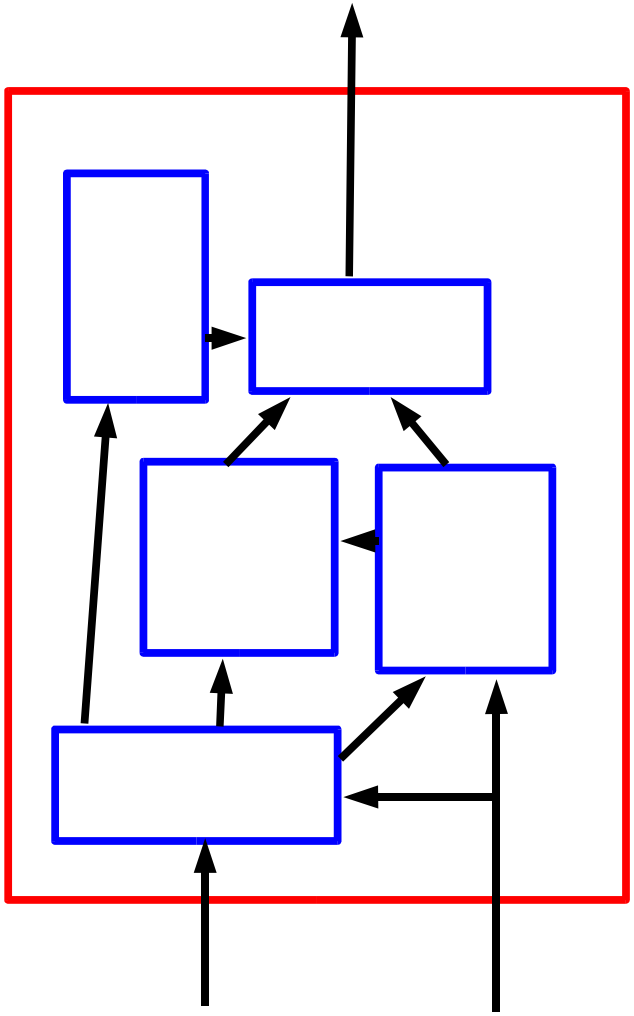
```
(defmethod nn-2layer fprop (input output)
  (==> layer1 fprop input hidden)
  (==> layer2 fprop hidden output) ())

(defmethod nn-2layer bprop (input output)
  (==> layer2 bprop hidden output)
  (==> layer1 bprop input hidden) ())
```

Assembling the Network: training



- A training cycle:
- pick a sample (X^i, Y^i) from the training set.
- call fprop with (X^i, Y^i) and record the error
- call bprop with (X^i, Y^i)
- update all the weights using the gradients obtained above.
- with the implementation above, we would have to go through each and every module to update all the weights. In the future, we will see how to "pool" all the weights and other free parameters in a single vector so they can all be updated at once.

Any Architecture works



- **Any connection is permissible**
  - Networks with loops must be “unfolded in time”.
- **Any module is permissible**
  - As long as it is continuous and differentiable almost everywhere with respect to the parameters, and with respect to non-terminal inputs.